Doctoral dissertation

High-precision estimation method through deep learning considering data distribution

(Wave-making resistance estimation

considering the distribution of ship figure)

データ分布を考慮した深層学習による高精度推定法

（船型データの分布を考慮した造波抵抗推定）

Graduate School of Engineering,
Yokohama National University

Xin LI
李 欣

Advisory Professor: Tomoki HAMAGAMI

March 2020

Abstract

With the progress of deep learning research in recent years, applications using end-to-end learning are expanding in various industries and fields. However, there are many situations where it is difficult to directly perform end-to-end learning on real data such as design problems. In particular, it is known that the learner becomes unstable when the data sampling density is non-uniform or when am imbalance exists. In this study, to solve these problems, we aimed at the method of extracting low-dimensional features of three-dimensional shapes and eliminating bias by prioritized selective sampling. The former clarified a method for efficiently expressing the features of three-dimensional shapes by extracting base frequency components from changes in the hidden layer of DAE. The latter clarified a two-level priority sampling method that obtains a robust learning result from a biased data set. This method was applied to the hull shape design problem, and the estimation accuracy was improved compared to the conventional method, and an end-to-end system that directly regressed the wave-making resistance from the hull shape was successfully constructed.

あらまし

近年の深層学習研究の進歩に伴い，様々な産業・分野で end-to-end 学習を用いた応用が拡がっている。しかし，設計問題をはじめとする実データにおいては end-to-end の学習を直接行うことが困難な場面が多く存在する。特に，データのサンプリング密度が不均一である場合やインバランスが存在している場合，学習器が不安定になることが知られている。本研究では，これらの問題を解決するために，3 次元形状の低次特徴量の抽出法と優先的選択的サンプリングによる偏りの解消をめざした。前者により，DAE の中間層の変化から周波数基底成分を取り出すことで 3 次元形状の特徴を効率的に表現する方法を明らかにした。また後者により，偏りのあるデータ集合からロバストな学習結果を得る 2 段階の優先サンプリング手法を明らかにした。本手法を，船型形状設計問題に応用し，従来手法に比べて推定精度が上がり，船型形状から造波抵抗を直接回帰する end-to-end のシステムの構築に成功した。

Acknowledgments

First of all, I am deeply grateful to my supervisor Prof. Tomoki Hamagami, who gave me such a very fascinating topic about deep learning to solve the problem of wave-making resistance estimation, and helped me a lot when I was confused. At the same time, more importantly, I learned a lot from his sincere attitude to research and his humanity.

Moreover, I sincerely appreciate Prof. Ryuji Kohno, Prof. Hideki Ochiai, Prof. Keisuke Shima, Prof. Chika Sugimoto, and Prof. Masaya Nakata for their valuable comments during the degree review, which has exceedingly helped me in the process of improving this dissertation.

At the same time, I would also like to thank the past and present members of the Hamagami laboratory for their hard work and dedication to research, which often gave me energy.

Finally, I would like to thank the China Scholarship Council for supporting the three-year living expenses of my doctoral program and my family for all kinds of support.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

In the rapid development of machine learning[1], deep learning[2], and the popularity of big data technology, the intelligence and accuracy of data processing is becoming more and more important, especially for some production industry. For example, in the case of the shipbuilding industry, there is a problem with how to generate a regression curve for wave-making resistance[3] of the ship. Here, the wave-making resistance is an indicator to evaluate whether the built ships have a good shape for navigation. While the ship is sailing at high speeds, the lower wave-making resistance value means the higher energy-saving performance. Among the trade in the world, maritime transport by ship accounts more than 99% in the world logistics, especially in Japan, maritime transport by ship accounts for 99.7% of the total weight of all transported goods[4]. Energy-saving technology and high efficiency for ship navigation are greatly significant not only for reducing transportation costs but also for environmental protection and carbon dioxide reduction.

1.1.1 What is the wave-making resistance of ship

The energy cost of ship navigation is significantly affected by the resistance caused between the air and the sea surface when the ship is moving[5]. These

Figure 1.1: An example of a ship sailing in the water

resistances include:

1. Frictional resistance caused by the fluid and hull surface[6];
2. Viscosity resistance caused by the disturbance of fluid around the object[7];
3. Wave-making resistance owing to the generation of waves by the hull[3], Figure 1.2 shows the ship figure image of the hull.

Of these resistances, wave-making resistance varies significantly depending on the shape and speed of the ship; however, its analytical estimation is a difficult task. Therefore, the conventional method must rely on measurements based on experimental tests with water tanks.

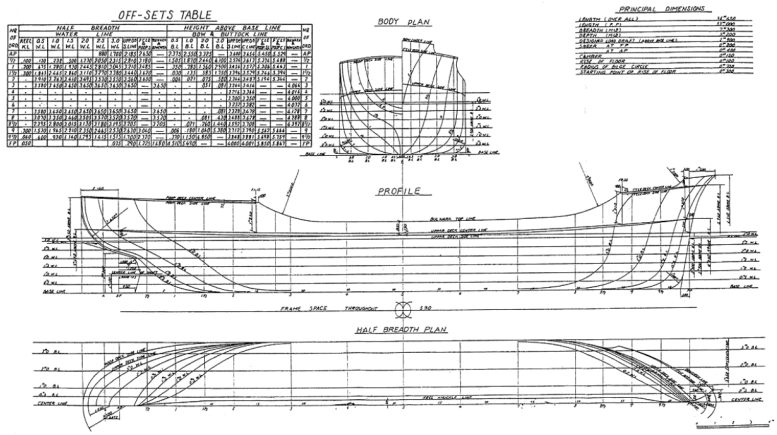


Figure 1.2: Ship figure image of the hull

1.1.2 Why it is necessary to change the wave-making resistance estimation method

As described in the previous section, a tank test requires large-scale equipment and time, the efficiency of the design is limited. It is better to perform the wave-making resistance estimation using an end-to-end learning algorithm[8], however, because of the ship parameters are not uniform, it is necessary to propose a set of solutions for processing such data, for it can be processed through deep learning. These solutions include interpolation, feature extraction, feature space conversion and so on.

1.1.3 About feature extraction

While processing data, dimensionality reduction[9] and feature extraction[10] are necessary works, also, if these works are not effective, they will greatly affect the accuracy of machine learning algorithms. Feature extraction is important for the whole process because of,

1. Using the extracted low-dimensional features as the input of deep learning helps to improve convergence speed and training efficiency.

2. A completed feature helps improve the accuracy of training.

Therefore, there are also necessary for the compressing of ship shape data.

As representative methods for the works, there are principal component analysis(PCA)[11] and auto-encoder(AE)[12]. PCA mainly solve linear problems, simultaneously AE mostly solve nonlinear problems. Because of AE has the structure of the neural network(NN)[13], it can be used in the expanding applicable scenes, as representative methods, there are AE, and its subspecies denoising auto-encoder(DAE)[14].

1.2 About this study

In this study, we will use the results of an already performed tank test and realize an end-to-end learning algorithm that directly obtains the wave-making resistance characteristic from the shape of the ship through deep learning. Moreover, while performing feature extraction for ship parameters, propose a feature extraction method that raises accuracy when data are skewed. By using these algorithms, we aim to support the automatic shape design[15][16] of hull design. Further, in the case of the same problem class, we look forward to solving other problems.

1.3 Overview of this dissertation

This dissertation is organized as follows. The following Chapter 2 described a method for wave-making resistance estimation from the shape of the ship through deep learning. To solve the problem in Chapter 2, Chapter 3 introduced an improved auto-encoder based on 2-level prioritized experience replay for dimension-

ality reduction of high dimension skewed data; Chapter 4 described a wave-making resistance estimation method considering the distribution of ship figure based on the method of Chapter 3; Chapter 5 discussed the conclusions and future works of this study.

Figure 1.3 shows the outline of this dissertation.

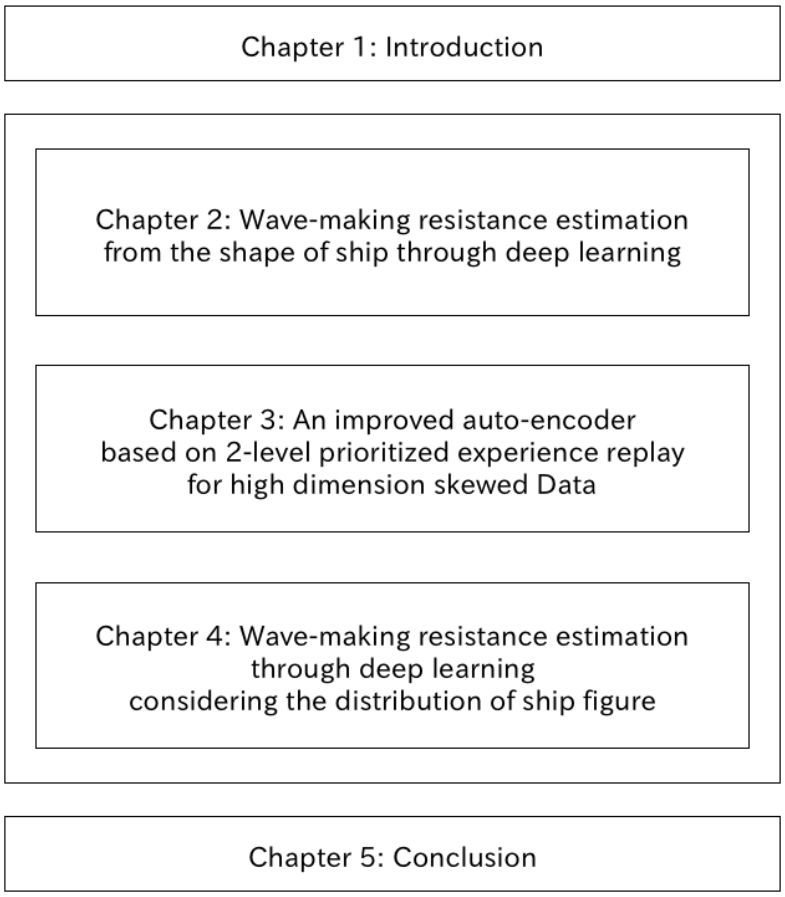


Figure 1.3: Outline of this dissertation

2 Wave-making resistance estimation from the shape of the ship through deep learning

2.1 Overview

A method for the estimation of wave-making resistance from the hull form and the Froude number through deep learning is proposed. The reduction of wave-making resistance is an essential issue in hull form design. However, the estimation of wave-making resistance is a time-consuming task that depends on experimental measurements. To enable direct estimation of the wave resistance from hull form, deep learning, which enables end-to-end learning, is an effective approach.

This chapter introduces a proposed method for the estimation of wave-making resistance, which has two phases.

1. First, auto-encoders, which reduce the dimension of the offset and the profile data are generated.

2. Subsequently, after the regularization of these data, a deep neural net for regression estimation of wave-making resistance is generated.

In Section 2.7 of this chapter there are the results of evaluation experiments. The results show that the proposed method can estimate wave-making resistance with high precision.

2.2 Conventional method

Conventionally, the following methods are using to estimate wave-making resistance.

1. Estimation based on actual measurement by the ship models

   Making ship models and performing water tank experiment, the experiments about measuring wave-making resistance parameter which does not depend on the scale has been widely performed[17]. Although the actual measurement experiment is the most reliable method, there is a limit to the efficiency of the hull design because of water tank measurement is required large-scale equipment and time. Figure 2.1 shows an image of actual measurement by the ship models.

2. Analytical method base on computational fluid dynamics (CFD)

   Computational fluid dynamics (CFD) as a method for obtaining wave-making resistance analytically from the shape of the ship is widely known[18]. By analyzing the motion of fluid generated around the hull dynamically, a simulation that not performed an actual measurement with the ship model is possible. Nevertheless, on the other hand, to set the computational grid and ensure stability, a lot of calculation time-consuming. Further, even if it changes a slight design on the ship, it requires an enormous amount of time. This is a remaining issue over the years. Figure 2.2 shows an image of CFD measurement for wave-making resistance.

For such an issue, this chapter will introduce a directly end-to-end estimation method for wave-making resistance from hull shape, which aims the same precision and estimation speed with the water tank experiment for wave-making resistance.

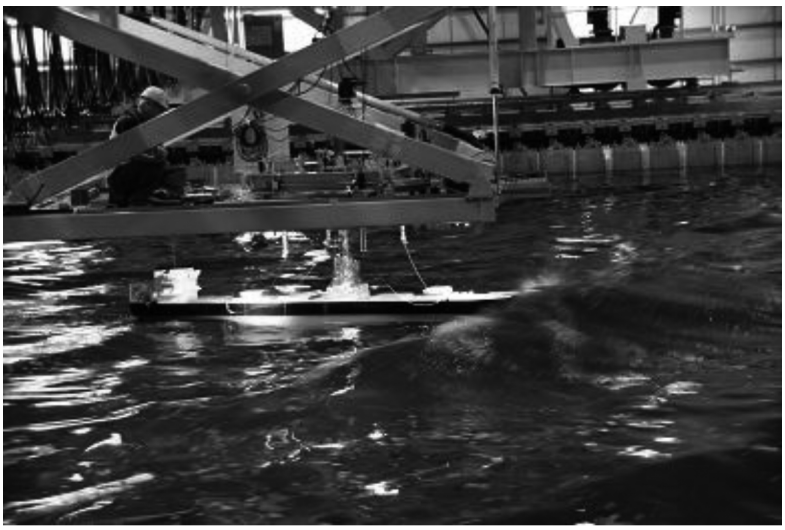


Figure 2.1: An image of actual measurement by the ship models

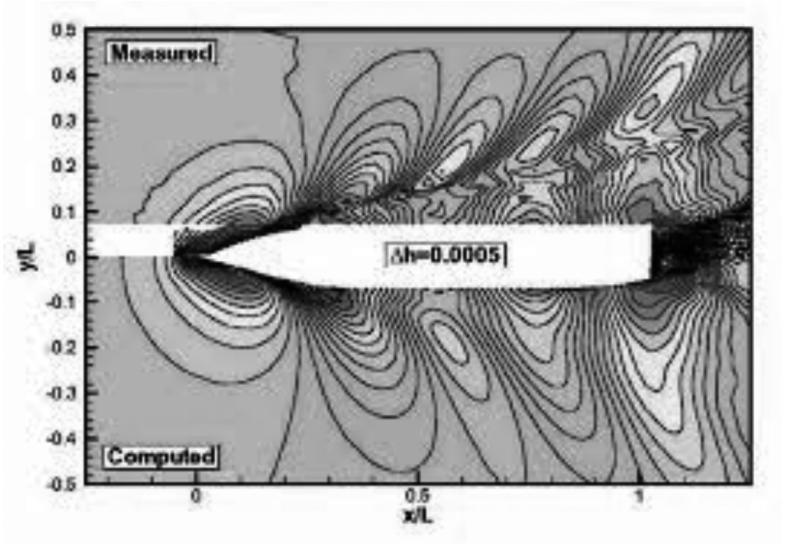


Figure 2.2: An image of CFD measurement for wave-making resistance

Specifically, the measurement experiment results of the water tank in the past will be used as training data, and a method through deep learning which can directly determine the wave-making resistance characteristic curve from the shape parameter of the ship will be achieved. Before introducing the proposed method, in the next section, the physical and mathematical definitions of wave-making resistance will be introduced specifically.

2.3 Wave-making resistance

As the introduction of Section 1.1.1 in Chapter 1, wave-making resistance varies greatly and can affect surface ship very seriously when the ships have different shapes and speeds. Ships need more energy to make water keep away from the hull of ships, and at the same time the energy creat waves. The specific principle is like this, when a ship sails, the surface pressure of the fore on the direction of motion increases, and the pressure creates waves on the surface of the water. Waves are generated by the pressure from the confluence of water divided into both sides of the ship at the aft. Owing to the interference of these waves, the complex energy loss of waves occurs and it becomes wave-making resistance. Wave-making resistance is determined by the speed of the hull and the velocity of the wave, in addition to the hull form. According to this phenomenon, it is known that wave generation is similar when the shape of the ship is similar. Therefore, the relationship between speed and wave-making resistance is expressed as a value obtained by nondimensionalizing the physical unit as follows:

$$F_n = \frac{U}{\sqrt{Lg}} \tag{2.1}$$

$$C_w = \frac{R_w}{\frac{1}{2}\rho U^2 S} \tag{2.2}$$

F_n represents Froude number and C_w represents the wave-making resistance coefficient. Here, U is the speed (m/s), L is the ship length (m), g is the gravitational acceleration (m/s^2), R_w is the wave-making resistance (N), ρ is the density (kg/m^3), and S is the hull inundation surface area (m^2). Both F_n and C_w are dimensionless parameters, and if the hull form is the same, the same relationship between F_n and C_w can be derived regardless of the scale. It is generally known that C_w increases as F_n increases. It is also known that the value of F_n, which shows an evident increase in C_w, and the trend of increase strongly depends on the hull form. In the design of hull shape, according to the F_n-C_w estimation, it is necessary to determine the hull shape when wave-making resistance most reduced in various scenes.

The current hull shape is examined by predicting C_w using computational fluid dynamics nowadays[19]. The performance of the ship type considered to be optimal for the examination of the hull form is confirmed using a water tank test. However, the theoretical calculation result and the experimental value are not necessarily the same because of the influence of various nonlinear factors. Therefore, the method which gets the relationship of F_n-C_w with a non-linear model through a direct regression is required.

However, the dimension of the parameters that define the three-dimensional shape of the various scales are not uniform, therefore, it is difficult to use these parameters to learning directly. Further, the data of the ship model used for regression has a large deviation in hull shape, there is a possibility that the learning results in bias. Therefore, the data pre-treatment processing is necessary, in the next section, the parameters of the original ship shape which used as training data for deep learning will be introduced.

2.4 The ship shape data

There are some differences in shape and size for each ship[20], but we have defined several general parameters for them to enable us to apply our proposed method. For tanker vessels according to the purpose of estimating wave-making resistance, data on vessels with different shapes, lengths, scales, etc. are targeted.

The shape of the ship is expressed by the following parameters.

- Offset (o_m)

  The first one is the offset part of the ship, which refers to the vertical cross-section shape against the length direction of the ship.

$$o_m = [w_1^m, w_2^m, \cdots, w_{N_o}^m] \tag{2.3}$$

  Here, $m = [1, 2, \cdots, M]$ is the cross-section number from the aft to the fore, $n = [1, 2, \cdots, N_o]$ is the number of samples in the cross-section m, w_n^m is the value of the contour sample n of the cross-section m, and represents the width from the centerline of the ship.

- Profile (p_f, p_a)

  The second one is the profile part of the ship, which refers to the vertical cross-section shape along the length direction of the ship.

$$p_f = [f_1, f_2, \cdots, f_{N_f}] \tag{2.4}$$

$$p_a = [a_1, a_2, \cdots, a_{N_a}] \tag{2.5}$$

  where p_f is the profile of the fore and p_a is the profile of the aft. N_f and N_a are the profile contour sample numbers of each profile, and $p_{[f,a]}$ is the sample value of each profile. The shape of the fore and aft on the centerline is expressed based on each reference point.

- Total length (L)

  The third one is the total length of the ship, which represents the length from the reference point of the fore to the reference point of the aft.

Figure 2.3 shows the relationship between the hull shape and each parameter[21].

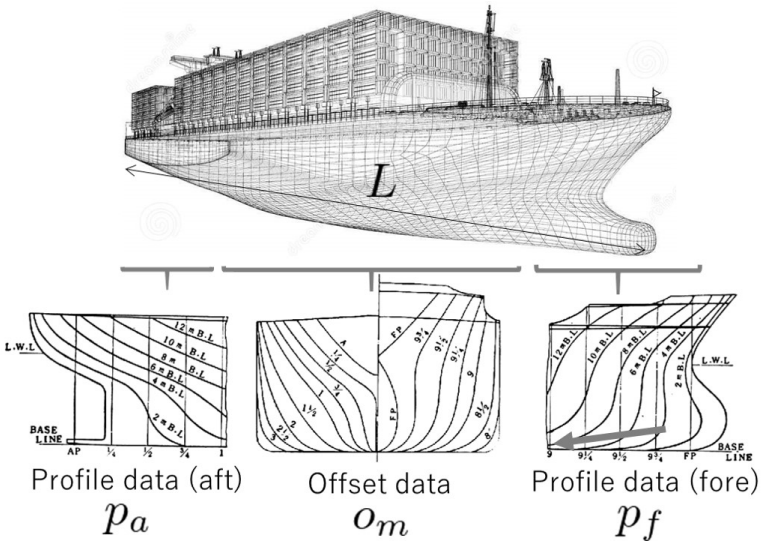


Figure 2.3: Ship plan and its offset and profile dataset

2.5 Pre-treatment for ship shape data

2.5.1 Interpolation and re-sampling

In order to generate the input data of next step for feature extraction, the ship shape data need to be interpolation and re-sampling. Therefore, for all o_m, p_f and p_a, perform interpolation operation and re-sampling.

2.5.2 Compression of shape parameter with auto-encoder

The variations in the shape of both the offset and profile are limited, and there is a possibility that the shapes can be expressed with lower-dimension features using the samples. Utilizing this property, we train an auto-encoder(AE), which expresses o_m, p_f, p_a in a low dimension, according to the following procedure. Incidentally, in the following process, the data set of all the ships is shuffled from o_m, p_f, p_a.

1. Use AE to learn the re-sample result. At this time, to increase the number of samples, we used denoising AE(DAE)[22] with Gaussian noise superimposed, for DAE, because of the Gaussian noise is added to the input so that not only the samples increased but also the general characteristics can be extracted for both training data and test data. Regarding the level of Gaussian noise, we used value with σ set to 3% of the data range for the input data. The reason for using Gaussian noise is that DAE needs a corruption distribution that adds noise to a portion of the input vector. As a corruption distribution, Gaussian noise or salt and pepper noise which randomly selects components and reduces them to 0 or 1 need to be used. However, to performing a maximum likelihood estimation, salt and pepper noise may destroy the data structure and Gaussian noise is a better choice.

2. Using the parameters of the intermediate layer obtained through learning, we obtained the dimension-compressed offset $\hat{o}_m$ and dimension-compressed profile $\hat{p}_f$, $\hat{p}_a$.

$$\hat{o}_m = [\hat{w}_1^m, \hat{w}_2^m, \cdots, \hat{w}_{\hat{N}_o}^m] \tag{2.6}$$

$$\hat{p}_f = [\hat{f}_1, \hat{f}_2, \cdots, \hat{f}_{\hat{N}_f}] \tag{2.7}$$

$$\hat{p}_a = [\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_{\hat{N}_a}] \tag{2.8}$$

Figure 2.4 shows a DAE used in the experiment described later.

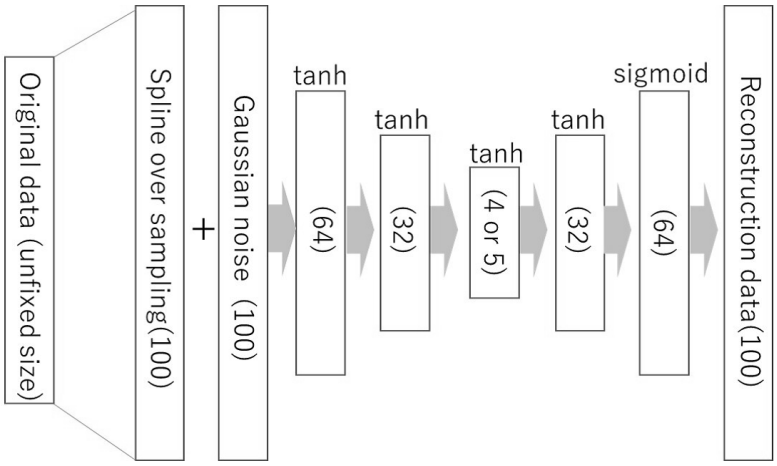


Figure 2.4: Denoising auto-encoder, which performs dimensional compression for ship shape parameters

As a preliminary experiment, we compared the reconstruction errors between PCA and AE while performing dimensional compression on the ship shape data. From Figure 2.5 we can know that for the feature extraction problem of ship shape data, a nonlinear method is better. And Figure 2.6 shows a comparison of the expression of the features between PCA and DAE.

When using DAE to perform feature extraction, the re-sampling number is set to 100 as neuron number, to determine the dimension of the intermediate layer, we performed a pre-experiment. Figure 2.7 shows the reconstruction error[23] changes evaluated by RMSE loss for different dimension of intermediate layer through cross-validation. In this figure, the solid line means average value, the broken line means variance ranges. Indeed, the higher the dimension, the closer to the original dimension number of the ship shape data, but the redundant information will also be increased. From the figure, we can know that for the dataset currently, the

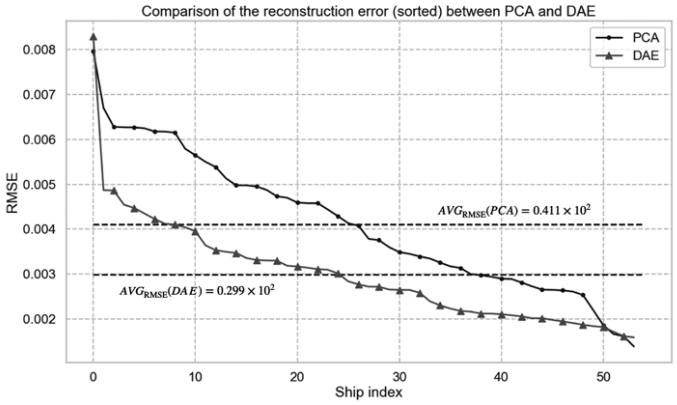
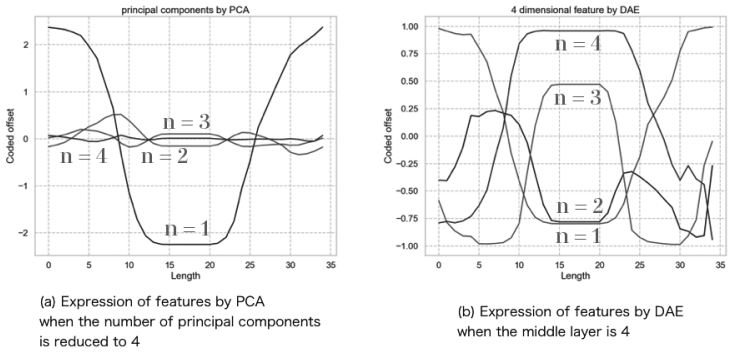


Figure 2.5: Comparison of the reconstruction error (sorted) between PCA and DAE



(a) Expression of features by PCA when the number of principal components is reduced to 4

(b) Expression of features by DAE when the middle layer is 4

Figure 2.6: Comparison of the expression of the features between PCA and DAE

number of intermediate layer nodes on 4 or 5 dimensions for DAE is a better choice as a feature dimension, which can ensure the learning gets the low reconstruction error.

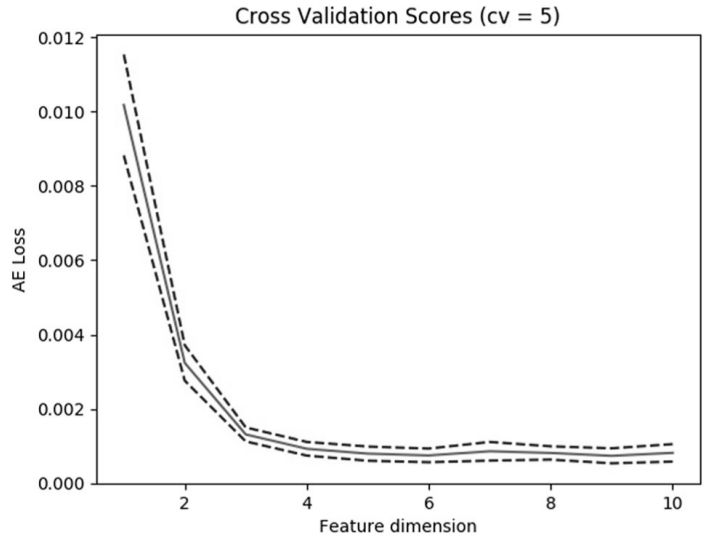


Figure 2.7: Changes in the reconstruction error owing to the difference in the DAE intermediate layer

Therefore, we set $\hat{N}_o = 4$, $\hat{N}_f = \hat{N}_a = 5$.

For the next step which performs a regression learning[24] in the next section, we will use the 5-dimensional features which we extracted through DAE.

2.6 Regression learning of the wave-making resistance

This section will introduce the method which directly regresses the relationship between F_n and C_w with a nonlinear model through end-to-end learning without physical calculation by using the parameters that define the shape of the ship. we

hope that through this we can accurately calculate the relationship between the F_n and C_w.

Using the compressed hull shape parameters $\hat{o}_m$, $\hat{p}_f$, $\hat{p}_a$ and any F_n as input, regression for C_w is performed by using a deep neural network(DNN).

The low-dimension compressed offset data obtained in the previous section have a different sampling number (M) in the length direction of the ship. Therefore, they cannot be added directly to the input layer of a fixed-dimension DNN.

To perform regression using DNN and do not generate significantly affecting on the hull shape, $\hat{o}_m$ is necessary to convert to a fixed-length input. Therefore, we have noticed that the even if the different ship hull, there are also common characteristics which $\hat{o}_m$ have no abrupt changes and always be smooth variation and continuous in the length direction. That is, present $\hat{O}_n$ as the series of each dimension of $\hat{o}_m$, when taken frequency components of $\hat{O}_n$ in the length direction, about the shape of the ship with continuous variation, because of the value of the continuous offset does not change suddenly, assume it can be approximated represent as a fixed low-dimension components of frequency.

2.6.1 Frequency components extraction

For specific instructions, generate the fixed-length feature of the frequency components according to the following procedure.

1. Decompose $\hat{o}_m$ into a vector $\hat{O}_n$ on each dimension.

$$\hat{O}_n = [\hat{o}_1^n, \hat{o}_2^n, \cdots, \hat{o}_M^n], \quad n = 1, 2 \cdots \hat{N}_o \tag{2.9}$$

2. For the sequence in length direction, perform zero padding and obtain a fixed length.

3. Convert it into complex frequencies using discrete Fourier transform(DFT) to analyze the frequency components, which include phase information of the sequence in the window of the ship length + zero paddings.

4. Among the obtained complex frequency components, until the fixed-length component which from a lower dimension to a proper dimension is referred to as a feature.

Since we want to preserve the ship shape characteristics of the hull, we do not need a window function which mainly processing to some normal signal. And about the phase information, it is important for securing the scale and length of the hull form. The innovative point of this method is that the shape data of variable length can be stored uniformly by frequency components.

Figure 2.8 shows an example of ship feature of different ships in length direction, and Figure 2.9 shows the process of DFT to convert features to fixed frequency components.

And Figure 2.10 shows the changes in the average reconstruction errors of the low dimensional features at each cut-off value of all ship shape data. To apply to all ship types, 20 is set with a margin since the average RMSE error (0.097) converges. Figure 2.11 shows the example of the offset feature when the cut-off be 20.

Through this operation, a fixed dimension of the offset can be obtained. The learning process of regression below will only use low-dimension components of frequency. Moreover, because of $\hat{p}_f$, $\hat{p}_a$ is fixed by the dimension of the intermediate layer of AE, it can be used to the feature for learning and do not need to convert to frequency.

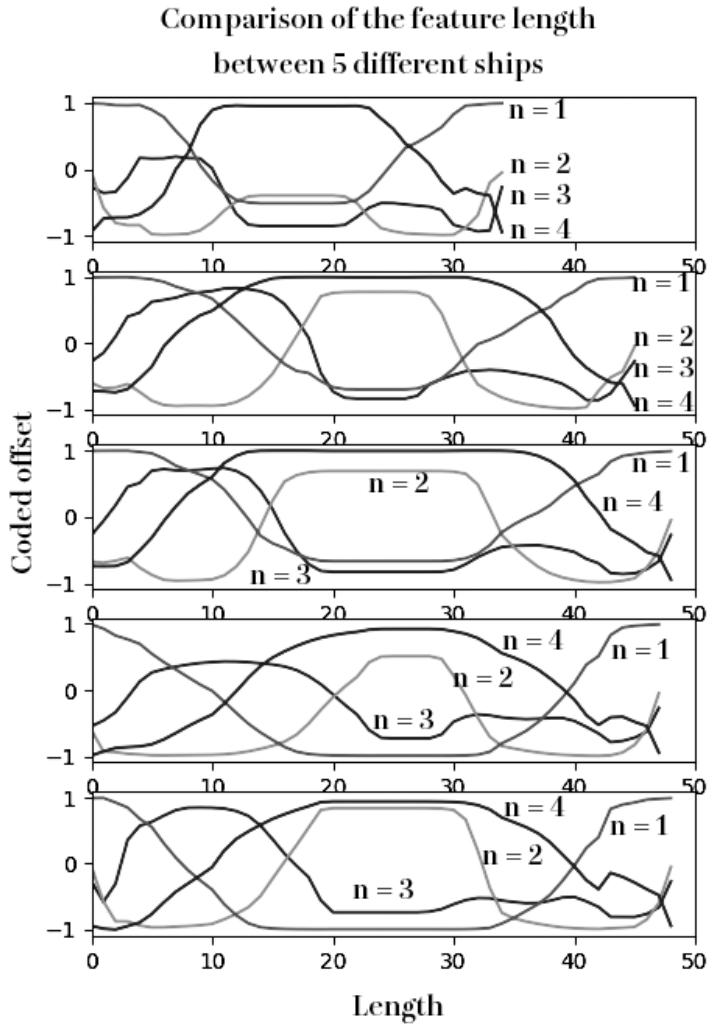


Figure 2.8: The offset feature sequence in different length

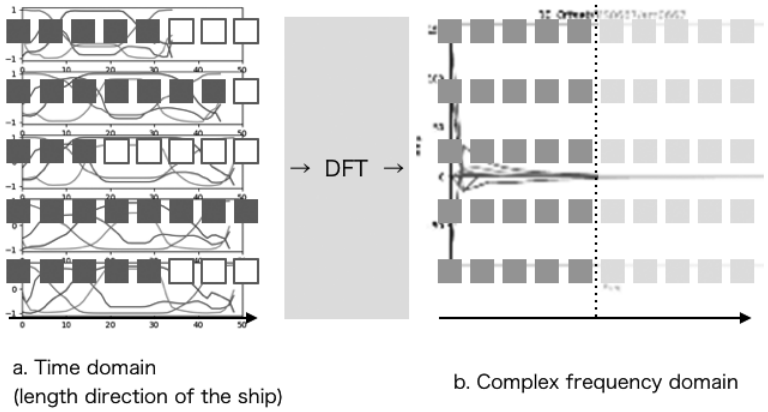


Figure 2.9: The image for DFT process

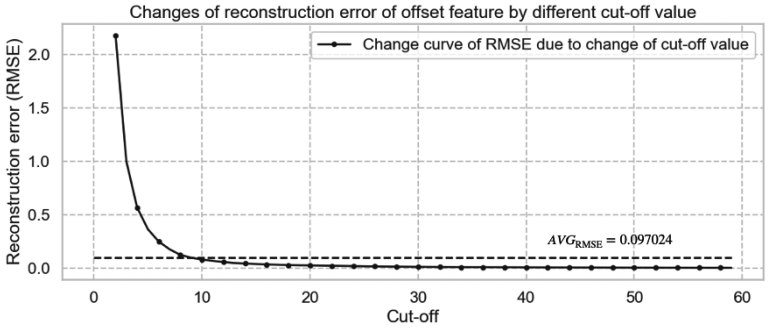


Figure 2.10: Changes of reconstruction error of offset feature by different cut-off value

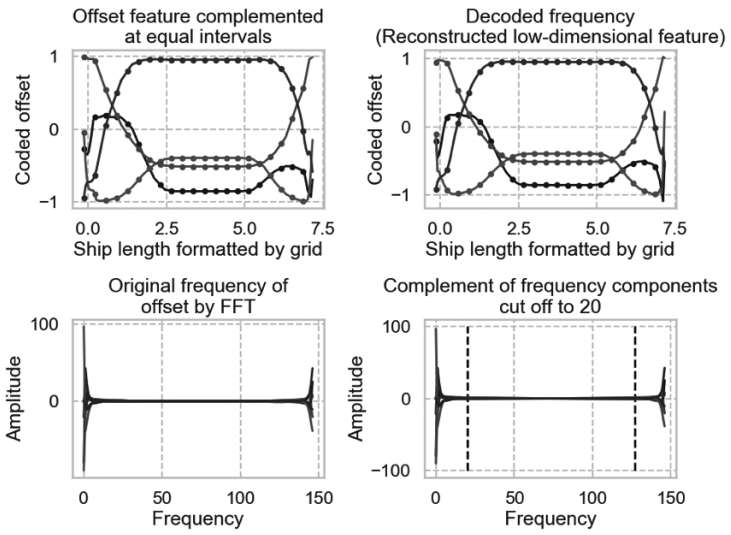


Figure 2.11: The features and the frequency components when cut-off be 20

2.6.2 Regression learning through deep neural network

Figure 2.12 shows a deep neural network for F_n-C_w regression used in experiment.

The overall view of the above procedure is shown in Figure 2.13.

2.7 Experiment

Using the above method and the data obtained from the water tank experiments, we conducted learning and evaluation experiments.

2.7.1 Experiment environment

Using the F_n-C_w obtained from the hull form parameters and water tank experiments of 58 model ships, we estimated the wave-making resistance via the proposed method.

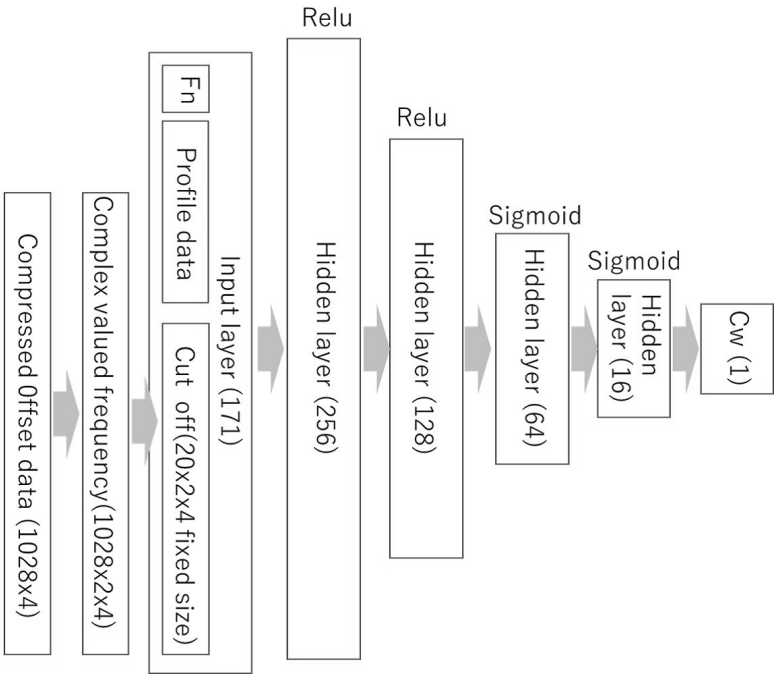


Figure 2.12: Deep neural network for F_n-C_w regression learning

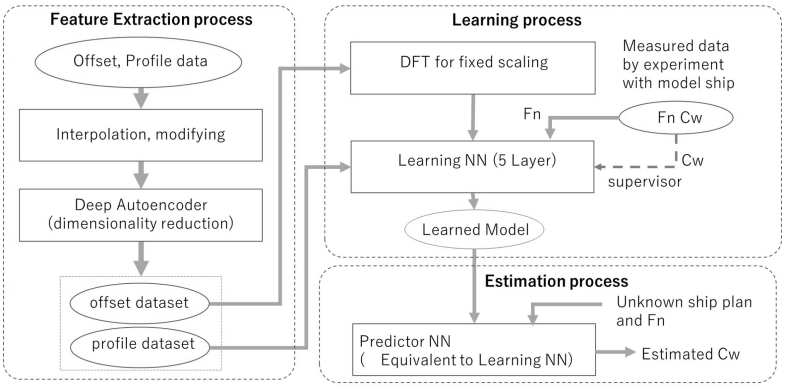


Figure 2.13: Procedure for estimating wave-making resistance from ship shape

Table 2.1 shows the specifications of DAE and DNN.

Table 2.1: Parameters of DAE and DNN for regression of F_n-C_w

Profile number	112×100
Offset number	2046×100
Cross-section number	$M \in [34, 49]$
Cross-section size	$N_o = N_f = N_a = 100$
DAE epoch number	100
DAE batch size	20
DAE layer	(100)-(64 tanh)-(32 tanh)-(4 / 5 tanh)-(32 tanh)-(64 sigmoid)-(100)
DNN epoch number	10000
DNN batch number	250
DNN input number	$171 = 5 + 5 + 20 \times 2 \times 4 + 1$
DNN layer	(171)-(256 relu)-(128 relu)-(64 sigmoid)-(16 sigmoid)-(1)

The adjustment of each hyperparameter of DAE and DNN was determined through preliminary experiments. Here, about the adjustment for the activation method, other than the output layer, hyperbolic tangent function (tanh) was used, and the output layer used sigmoid function (sigmoid). From the Figure 2.14 and Figure 2.15, we can know that tanh can guarantee that learning will not lose valid minus values in the hidden layers, and for the output, we need the values to revert to the range of 0 to 1 to correspond to the normalized ship shape data.

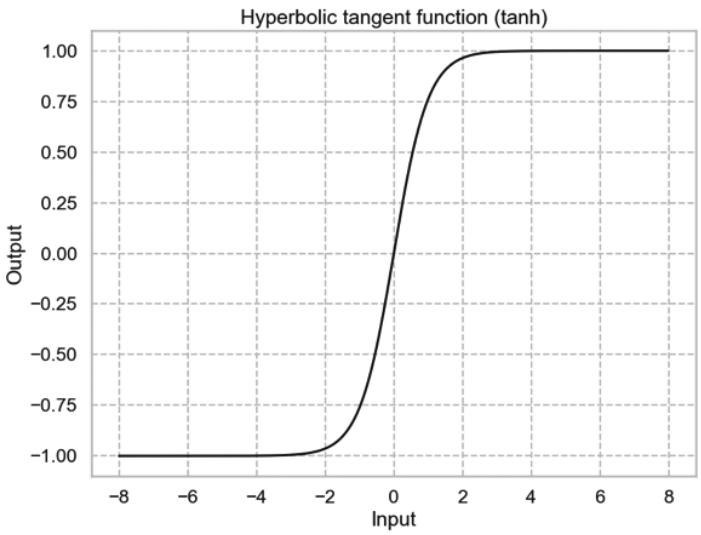


Figure 2.14: The hyperbolic tangent function

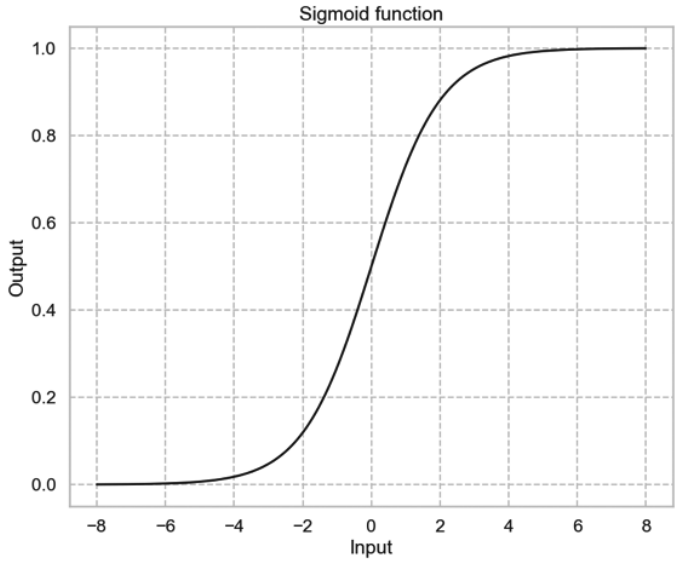


Figure 2.15: The sigmoid function

2.7.2 Experiment result

Figure 2.16 shows an example of the input and output from the learned DAE obtained for a certain ship shape. We can confirm that, for both offset and profile, the original shape can be reproduced. Moreover, for the offset change(offset parameter) in the length direction of the ship, although the change at the fore and aft is large, there is almost no change in the ship's body part, and regarding the expression of the offset frequency component, it is understood that the values are concentrated in the low dimension.

Appearance frequency distribution of RMSE error when the intermediate layer of DAE is four-dimension which we performed in the experiment is shown in Figure 2.17. Through the figure, we can know that the curve shows the approximate distribution of the data which following a chi-squared distribution. Most of the errors are relatively small, about the worst cases, a few of the data with large errors are shown in the figure which is considered to perform sampling to increase accuracy.

Based on these features, we used the DNN to perform regression learning for F_n-C_w (Figure 2.18). We evaluated the error in C_w for all ship types and the average RMSE was 0.078. Compare to the RMSE error 0.1 which normally gets from conventional method DFT, and it can be observed that the proposed method has been estimated with high accuracy, which raised an accuracy in 22% on wave-making resistance estimation. Figure 8 shows the results of sorted C_w errors in the descending order of F_n-C_w regression for all ship types. According to the experiment for the 58 ships, about the wave-making resistance estimation precision, the reconstruction errors for most of the ships were less than the average, however, minority ships are begetting the high errors. It can be observed that sufficient accuracy has not been obtained for C_w estimation of some types of ships.

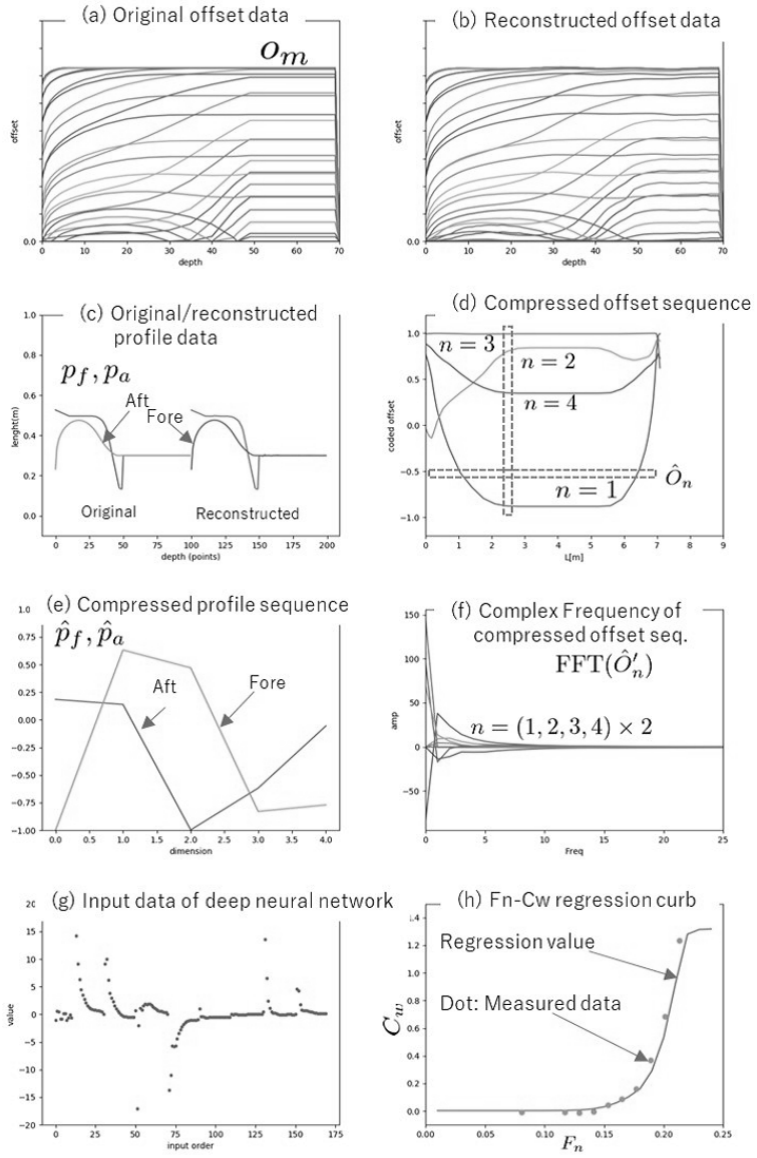


Figure 2.16: Reconstruction result of offset and profile, intermediate expression, and DAE regression result

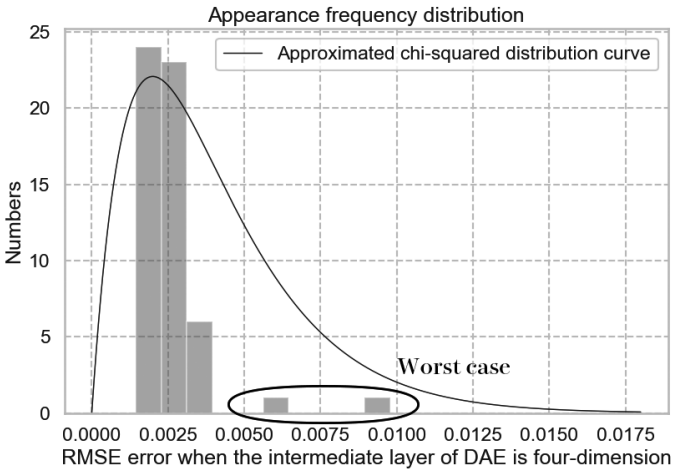


Figure 2.17: Appearance frequency distribution of RMSE error

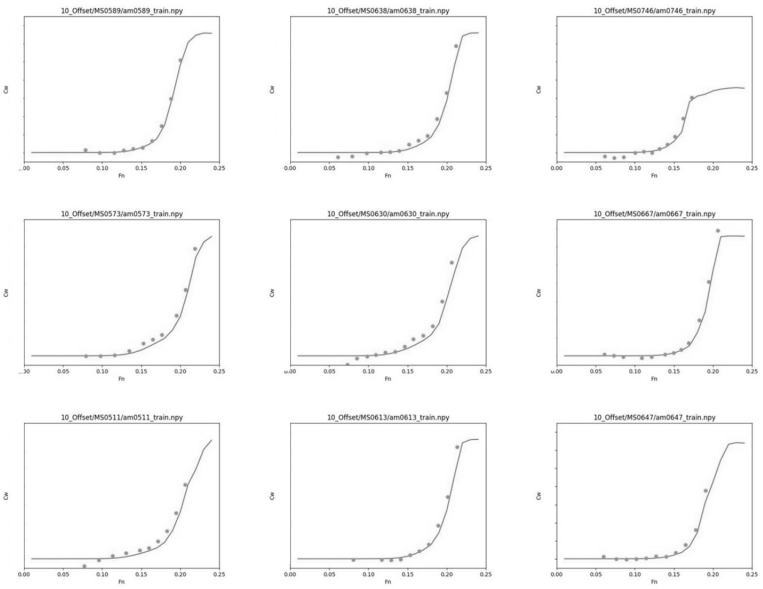


Figure 2.18: Result of regression analysis

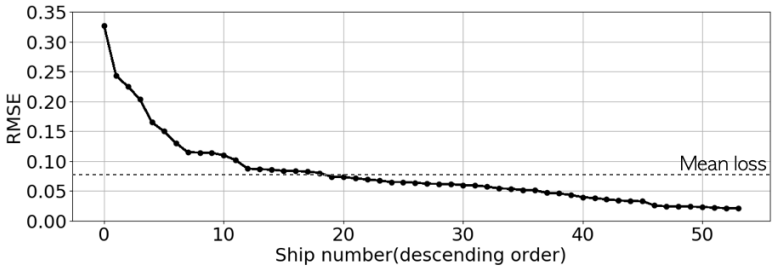


Figure 2.19: Regression error distribution evaluated through RMSE (descending order)

2.7.3 Consideration

The cross-section shape of the ship over a wide range of the length direction is rectangular and we refer to it as the parallel portion. Therefore, DAE learning deviated from the rectangular part. Relatively, the shape near the fore and aft may not be accurately expressed. It is considered that this error leads to the estimation error of F_n-C_w.

Therefore, we speculate that if we use some method that can conduct targeted and effective learning on the offset data for the fore and aft part, the error rate may be able to be reduced. Such a method needs to reduce the DAE reconstruction error of offset data for the fore and aft part while ensuring the low error of most of the original parts. To perform an investigation, a comparative experiment was conducted on this. From Figure 2.20, we find that the top and bottom are different ships. The first row is a plot of the offset in the ship's direction, and the second row is the form of the offset compressed in the lower order. In the third column, the blue curve is the regression curve and the orange point is the actual measurement. As you can see from the regression curve, the upper hull has higher reproducibility than the lower hull. The reason why the hull form below could not be reproduced well is that the characteristics of the offset aft and fore area of the ship are very

small in number from the overall data. It can also be seen that the features of the corresponding parts in the low-order region are not well extracted. Therefore, the effect of data bias is considered. This problem was solved by the general-purpose feature extraction method proposed in Chapter 3.

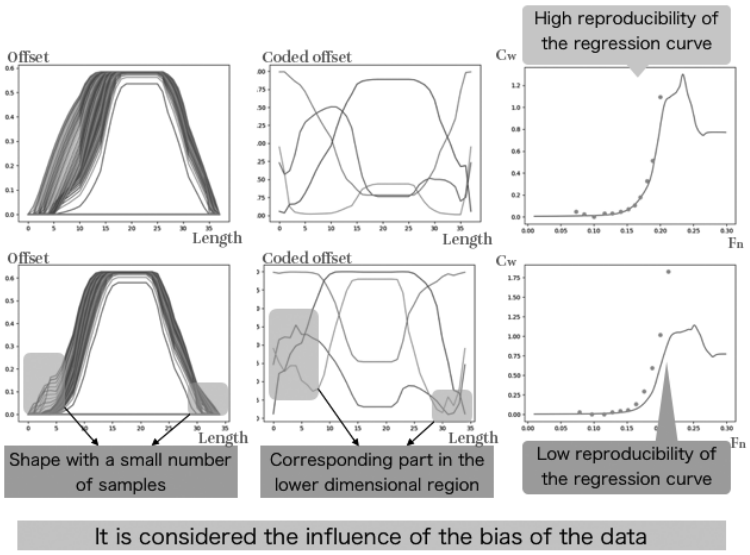


Figure 2.20: The comparison between two ships (one ship has the majority features and one ship have the skewed features)

2.8 Summary

We proposed a deep learning method to obtain wave-making resistance from the ship shape through end-to-end learning.

In the learning process, low-dimensional compression of shape parameters was performed using DAE and a frequency component in the length direction was used for obtaining fixed-length data. The results of the experiment revealed that, for

many types of ships, wave-making resistance characteristics can be obtained with high accuracy.

About how to absorb modality differences, Firstly, when the number of points in each section of the hull form is not uniform, a method of zero padding to make them the same number of dimensions, and this method is applicable when the number of points in the scale-independent shape data is not uniform; Secondly, extracting fixed components of frequency components when it is not possible to directly input to the input layer of deep learning because the length of the data is not uniform. This method is applicable to shape data that has the characteristic of changing smoothly.

In the next Chapter 3, we will introduce a method that considered the bias of dataset, and we expect that through the method could improve the proposed method of this chapter for raising the overall learning accuracy of offset data.

3 An improved auto-encoder based on 2-level prioritized experience replay for high dimension skewed data

3.1 Overview

Auto-encoder(AE) has been used for the feature extraction in Chapter 2 while preparing the input data for deep regression. However, because of the bias of the training data, the proposed method on Chapter 2 did not obtain high accuracy on every ship. In this chapter, AE will be introduced in detail and based on the AE and DAE, we proposed an improved feature extraction method that considering the data distribution.

AE as the representative method for data dimensionality reduction and feature extraction plays a very important role in machine learning. However, the data in the actual research work of institutions or industrial production is not always normalized. Moreover, regarding real data, there are many data are skewed in actual production and life. Most of the production which involves the shape data, when considering the optimal solution for a particular standard, the shapes are often irregular and bring about data skew, at this time, it will lead to high reconstruction error and slow convergence speed.

For example in the case of shipbuilding which we introduced in Chapter 2, the length between perpendiculars (LBP)[25] of ships always be huge, meanwhile, the range of aft and fore of the ship is always short. If observing the whole data of the ship, we can know that the data of aft and fore of the ship are skewed, which have a low percentage on the amount of data and follow a different distribution, at this time the reconstruction error[23] of compressed data or the extracted feature will be high due to the influence of the skewed data, in this case, the learning process will be hard to work.

For this problem, any variant of AE does not reduce the error very well, at this time, although the shipshape data are skewed, precise accuracy and convergence speed are still relevant. The accuracy of algorithms directly affects productivity, therefore, based on these legacy issues, in this chapter, we aim at a feature extraction method, which can get low loss, and fast convergence speed for high dimension skewed data.

This chapter will introduce an improved auto-encoder and a denoising auto-encoder based on 2-level prioritized experience replay, which can improve accuracy and reduce loss, while processing a dimensionality reduction or feature extraction problem on high dimension skewed data. To evaluate the effectiveness of the proposed method, three models of high dimension simulation dataset which on different skewed degrees are generated. The results of evaluation experiments show that the proposed method can get lower reconstruction error than the conventional method for high dimension skewed simulation data.

Compared with other approaches to skewed problems, the first focus of this study is to improve the accuracy of feature extraction for skewed data. Other approaches have solved skewed problems when the data are skewed as the prior knowledge, in this study, we considered to solve the problem through optimize the

learning process[26].

In the following, Section 3.2 describes the original AE and DAE; Section 3.3 describes how to reduce the loss rate by using the proposed method with sampling; Section 3.5 shows the experimental results through a comparison between the conventional method with the proposed method on high dimension skewed simulation data; Section 3.6 is a summary of this chapter.

3.2 Introduction of AE

Auto-encoder(AE) is an unsupervised learning[27] algorithm mainly used for data dimensionality reduction and feature extraction.

In deep learning, AE can be used to determine the initial value of the weight matrix before the training phase begins. The weight matrix in the NN can be regarded as the feature conversion of the input data, that is, the data firstly encoded into another expression, moreover, then a series of learning performed base on this.

However, when initializing the weights, we don't know what role of initial weights plays in training, nor how the weights will change during the training. Therefore, a better idea is that when encoding with the weight matrix generated by initialization, we hope that the encoded data can better retain the main features of the original data.

About how to measure whether the encoded data retain complete information, from previous research, if the encoded data can be easily restored to the original data by decoding, in other words, the lower reconstruction error, the features of the data are preserved well.

AE has such an NN structure, which mainly constituted by an encoder and

a decoder, and perform backpropagation to solve a nonlinear reconstruction optimization problem. Figure 3.1 shows the structure, and equation 3.1 shows a function definition of AE.

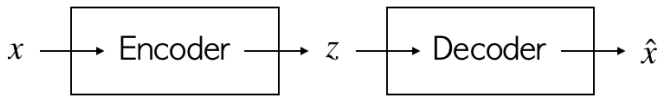


Figure 3.1: The structure of AE

$$\hat{x} = f(x; \theta) \tag{3.1}$$

Here, x is input data, and $\hat{x}$ is reconstruction data. z is an under-complete representation generated from the intermediate layer, which contains a feature of input data. θ is a weight matrix that contains parameters of AE. Through calculating the reconstruction error $L(x, \hat{x})$ between x and $\hat{x}$, the effectiveness of AE can be evaluated.

$$L(x, \hat{x}) = \sqrt{||x - f(x; \theta)||^2} \tag{3.2}$$

$$\theta = \arg\min \frac{1}{N} \sum_{n=1}^{N} L(x_n, \hat{x}_n) \tag{3.3}$$

Equation 3.2 shows the definition of error function, and equation 3.3 shows the target θ which we want to get, where data number is $n \in 1 \cdots N$.

AE also have some variants, to take a simple example, it contains denoising auto-encoder(DAE), variational auto-encoder(VAE)[28], or beta-VAE[29] etc.

DAE mainly used to generate generalized features by adding noise to input. VAE mainly used to get a generation model which can control output through

adjust latent variable. Moreover, if a disentangled representation, which differs from the distributed representation and mapping one is a hidden variable to one meaning, is necessary beta-VAE is a better choice.

However, the common problem with these methods is that, when dealing with the problem with high dimension skewed datasets, they can not reduce the error very well, moreover, the learning results are biased. That is, the problem of a decrease in generalization is shown.

3.3 Proposed method

To solve the problems described in Chapter 2, in this chapter, a sampling method for AE and DAE on the high dimension skewed dataset is proposed.

Fig. 3.2 shows the AE, and Fig. 3.3 shows the DAE used in the chapter of the experiment.

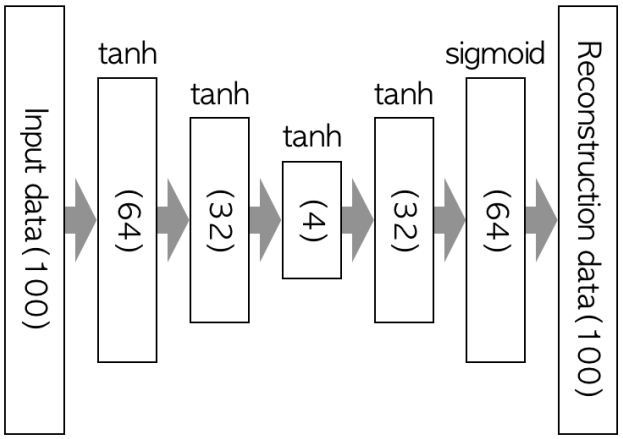


Figure 3.2: The AE structure used in the experiments

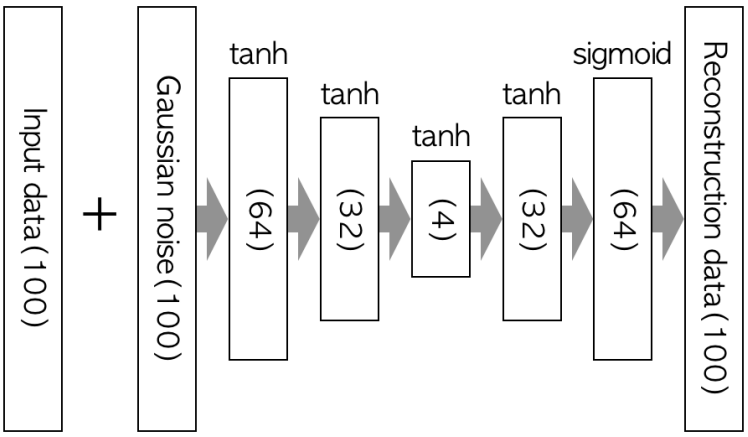


Figure 3.3: The DAE structure used in the experiments

Here when using DAE, the noise will be added by the distribution of Gaussian noise. Figure 3.4 shows the distribution of Gaussian noise.

From previous research, in recent years, prioritized experience replay(PER)[30] which can enhance generalization performance through preferentially reproducing experiences on the data with few opportunities, is proposed using reinforcement learning. Figure 3.5 shows the reinforcement learning solutions by prioritized experience replay for unstable learning.

From examining the findings, in this research, we use priority sampling[31] to enhance opportunities for high dimension skewed data, moreover, enhance generalization performance for AE and DAE.

Equation 3.4 shows the definition of PER.

$$\mathrm{PER}_i(\alpha) = \frac{p_i^\alpha}{\sum_k p_k^\alpha} \tag{3.4}$$

Here, p_i is the reference sampling probability based on error for sample i, and

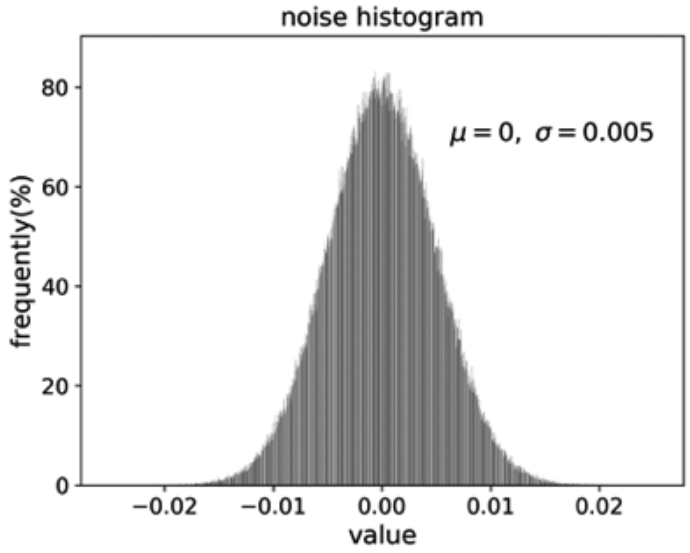


Figure 3.4: The distribution of Gaussian noise which used for DAE

this probability can be used to perform a roulette selection. $\mathrm{PER}_i(\alpha)$ is selection probability for sample i, α is weight parameter, within the range of $0 < \alpha < 1$, $\mathrm{PER}_i(1.0) = p(i)$, and $\mathrm{PER}_i(0.0)$ makes the learning become a random sampling.

In this research, considering the idea that using samples with significant errors preferentially for learning, we are setting the reference sampling probability according to the reconstruction error of AE and DAE.

For example, (1) firstly, starting learning progress from random sampling($\alpha = 0$); (2) secondly, according to learning progress and variation of reconstruction error, update α; (3) finally, taking $\alpha = 1.0$ and finish learning progress.

However, when PER is performed for all samples, the sampling priority strongly depends on the distribution of reconstruction error.

For example, even if there are a small number of samples with significant reconstruction errors, if there are a large number of samples with small reconstruction errors, there is a possibility that the sampling does not work.

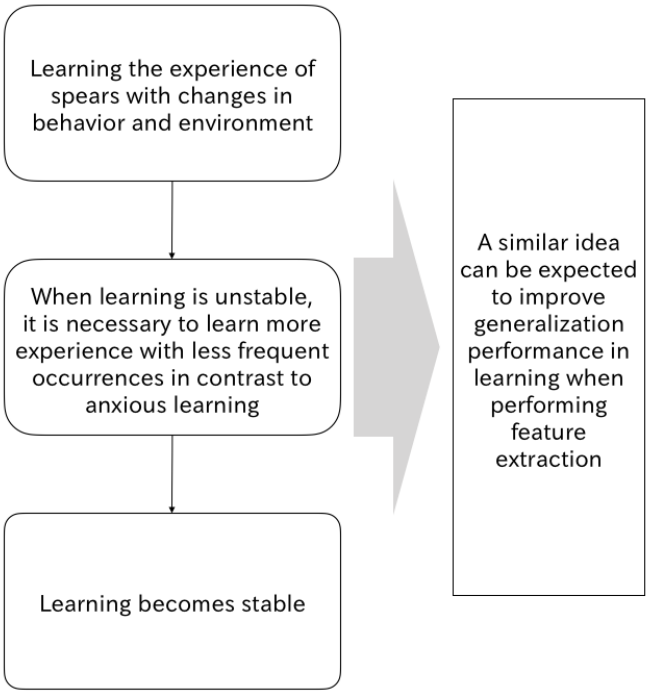


Figure 3.5: The reinforcement learning solutions by prioritized experience replay for unstable learning

Here, we set an arithmetic progression with an element number of 100 as an example of the probability array, the sum of all elements in the array is 1.0, Fig. 3.6 shows the relationship changes between $p(i)$ with $\mathrm{PER}_i(\alpha)$ on the different α settings for this array. It can be known from this that even in the case of 100 data if the sample of high-error data is tiny, the effect of PER will be weakened.

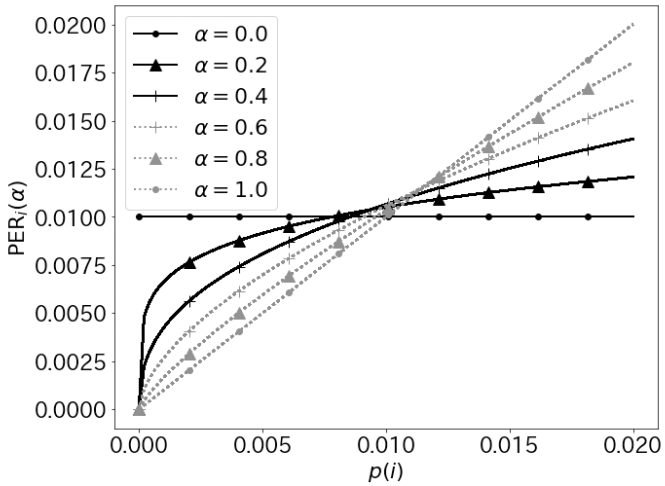


Figure 3.6: The relationship changes between $p(i)$ with $\mathrm{PER}_i(\alpha)$ on the different α settings for an example of probability array

Moreover, the problem is that the actual application does not have only 100 data, usually with larger samples. Therefore, we proposed a method based on 2-level PER sampling, and make samples with significant reconstruction error can be preferentially selected anytime without depending on the sample number. Figure 3.7 shows an improved AE or DAE based on 2-level PER, and the procedure of the proposed method is shown below.

1. Let $\alpha = 0$ in first iteration, and learn to a EP epochs. When finish learning,

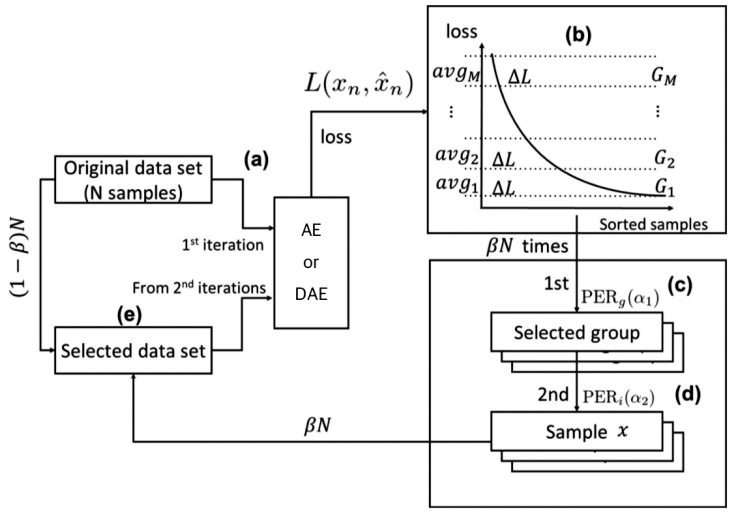


Figure 3.7: Improved AE or DAE based on 2-level PER

using the obtained AE / DAE, calculate the reconstruction error $L(x_n, \hat{x}_n)$ for all data $n = 1 \cdots N$. (Fig. 3.7(a))

2. Divide $L(x_n, \hat{x}_n)$ into groups $m = 1 \cdots M$ in units of ΔL, calculate the average value avg_m of reconstruction error for each group. (Fig. 3.7(b))

3. Let avg_m as reference probability, select a group based on $\mathrm{PER}_g(\alpha_1)$. (Fig. 3.7(c))

4. For samples belonging to the selected group, according to the reconstruction error of these samples, select a sample based on $\mathrm{PER}_i(\alpha_2)$. (Fig. 3.7(d))

5. Extract βN samples as described above, and select $(1-\beta)N$ samples randomly from all samples. (Fig. 3.7(e))

6. Use the selected samples as training data and learn to EP epochs.

7. Go back to 2.

Figure 3.8 shows an image of the comparison between not using sampling and using sampling.

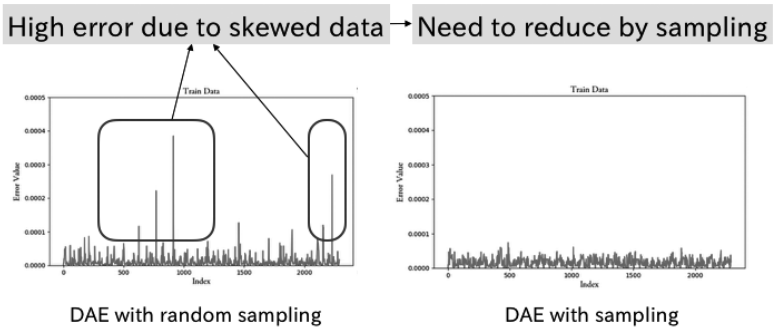


Figure 3.8: A comparison of reconstruction error through DAE between not using sampling and using sampling

Here, the x-axis means the index of every sample, and the y-axis means the error of each sample. We can find the error of high error samples will be reduced by sampling.

3.4 Dataset

To evaluate the effectiveness of the proposed method, we generated three models of high dimension simulation dataset which on different skewed degrees.

For example,

1. Firstly, generate two different normal distributions, according to distribution 1 and distribution 2, let μ_1, μ_2 be their expectation, and σ_1, σ_2 be their standard deviation;

2. Secondly, generate 95% samples from distribution 1, moreover, generate 5% samples from distribution 2 using as skewed data;

3. Finally, generate d dimension samples as described above.

Table 3.1 shows the settings for simulation data.

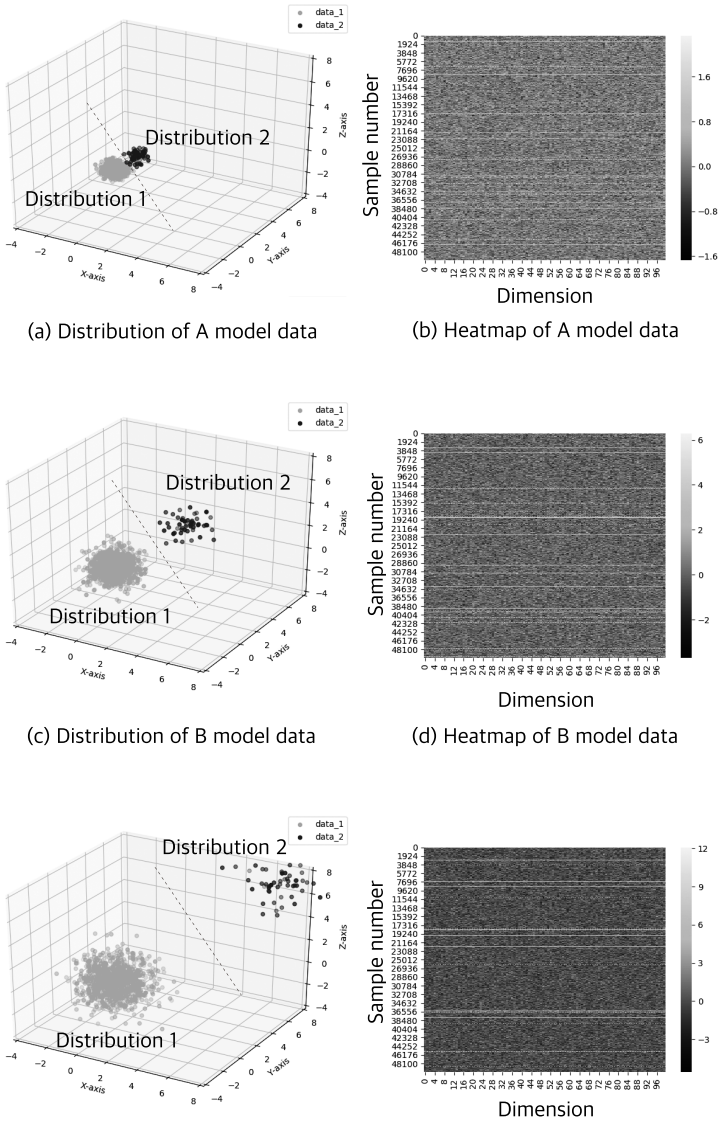
Table 3.1: The skewed degree control for three models of simulation data

	A model	B model	C model
μ_1	0.0	0.0	0.0
μ_2	1.0	3.0	7.0
σ_1, σ_2	0.1	0.5	1.2

In the experiment, we generated 50000 samples of dimension $d = 100$. To control the correlation between each dimension, the variance-covariance matrix is the right choice.

From previous research, when extracting features, it is more prone to skewed issues for the data with a low correlation between dimensions. Here we let the covariance to be 0 on every two dimensions of the three models of simulation data to reduce correlation.

Fig. 3.9 shows the three models simulation data, (a)(c)(e) show the three models of simulation data on three coordinate axes, however, the data have 100 dimensions, therefore considering any three dimensions in 100 dimensions data have similar relative relationships. (b)(d)(f) show the heatmap of three models simulation data, through which we can know the situation of skewed degree for the data of each model.



(a) Distribution of A model data

(b) Heatmap of A model data

(c) Distribution of B model data

(d) Heatmap of B model data

(e) Distribution of C model data

(f) Heatmap of C model data

Figure 3.9: The high dimension skewed simulation data of three models

3.5 Experiment

3.5.1 Experiment environment

We used Keras[32] as an experiment tool for each experiment below. Figure 3.10 shows the model structure in Keras of AE or DAE for training, and when DAE, the input is corrupted by Gaussian noise[33], with 3% of data range as σ to input data.

Table 3.2: The common parameter settings of improved AE and DAE

Simulation data number	50000
Training data number	36000
Validation data number	9000
Cross-validation data number	$45000 = 36000 + 9000$
Test data number	5000
Sampling group stride	$\Delta L = 0.06$
Sampling factor	$\beta = 0.2$
Sampling parameter	$\alpha_1 = 0.3, \alpha_2 = 0.5$
Sampling times	4
Epoch number for each smapling	$EP = 100$
Total epoch number	$400 = 4 \times 100$
Batch size	20
Layer	(100)-(64 tanh)-(32 tanh)-(4 tanh)-(32 tanh)-(64 sigmoid)-(100)
Loss function	RMSE(Root Mean Squared Error)

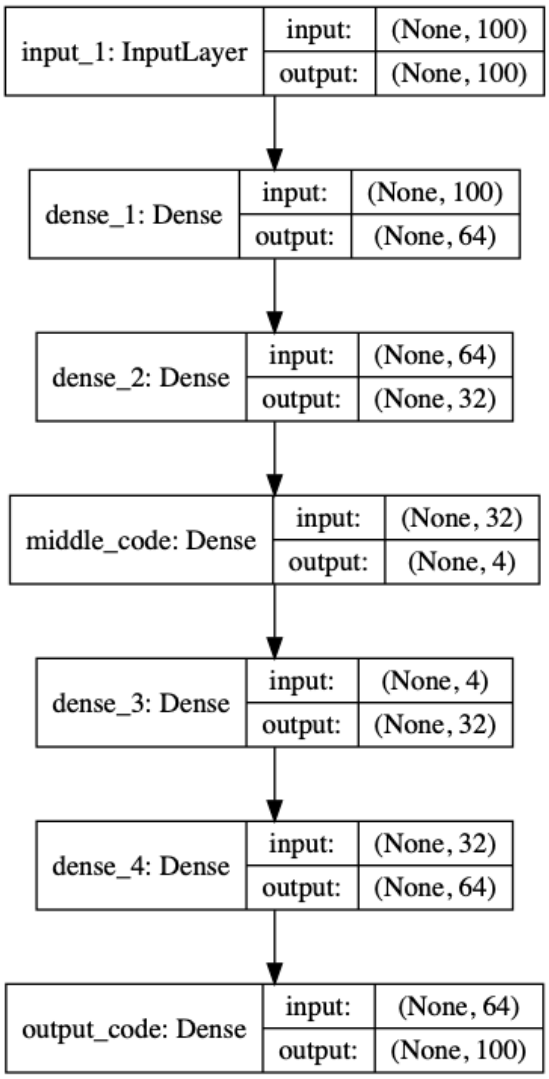


Figure 3.10: The model structure of AE or DAE for training

We use the values in Table 3.2 as common parameter settings for improved AE and DAE, the parameters were already adjusted to appropriate values through preliminary experiments. About the parameters adjustment for α_1 and α_2, in most cases, we hope the parameter values to be higher, however when the skewed degree increased, the learning process becomes difficult, if increase parameter values blindly, it may lead to the training data changes intensely during sampling, this will bring a lot of instability, it is usually embodied as the learning curve up and down cyclically, therefore, this is a trade-off relationship.

A feasible solution is to increase the value under the premise that learning can proceed normally, and the preliminary experiment was also carried out in this way. However the standard parameters are not always valid for all experiments, they need to be adjusted for some specific situations.

About the experiment in which the parameters need to be changed, we adjusted some values additionally. In this experiment, when the data of the C model has been used, Because of the distribution of the C model perhaps lead to the learning got high reconstruction error relatively, the sampling group stride must be adjusted to a new appropriate value, in general, because of the error range becomes wider, the sampling group stride also follows the larger adjustment.

Moreover, according to high reconstruction errors due to a highly skewed degree, to get more efficient features from learning, it also requires an increase in the sampling parameter through an adjustment on α_1 and α_2. Therefore, when the processing target is C model data, some of the changes in the parameter settings are shown in Table 3.3 and Table 3.4.

Here, the parameters settings are based on some preliminary experiments. Regarding the difference in adjustment for improved AE and improved DAE, we consider that is because the DAE already has some generalization performance

Table 3.3: The changes of parameter settings for improved AE on C model simulation data

Sampling group stride	$\Delta L = 0.09$
Sampling factor	$\beta = 0.3$
Sampling parameter	$\alpha_1 = 0.5, \alpha_2 = 0.7$

Table 3.4: The changes of parameter settings for improved DAE on C model simulation data

Sampling group stride	$\Delta L = 0.08$
Sampling factor	$\beta = 0.3$
Sampling parameter	$\alpha_1 = 0.3, \alpha_2 = 0.8$

due to the addition of Gaussian noise, therefore compare to the AE, the value of α_1 which controlling the group selection can be lower, about the value of α_2 can be almost the same level as the AE.

To compare the proposed method with AE and DAE, 5-fold cross-validation[34] has been performed. Fig. 3.11 shows the splits of samples through cross-validation on the experiment, here, for the simulation data of each model, 10% are generated for test data through random selection, which should be fixed and do not change. The cross-validation is performed on the remaining 90%, therefore, for 5-fold cross-validation, 72% training data and 18% validation data are generated for the cross-validation of each time, at the same time, we always have the 10% test data to perform an additional validation.

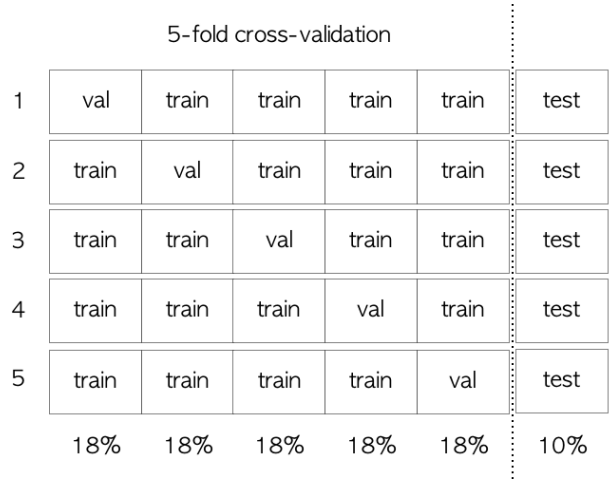


Figure 3.11: The splits of samples for A, B, C models simulation data through cross-validation

3.5.2 Experiment result

We performed AE, DAE, and the proposed method on A, B, C models simulation data by 5-fold cross-validation. Fig. 3.12 shows the comparison of the RMSE average by reconstruction error between AE and the proposed method, and Fig. 3.13 shows the comparison of the RMSE average by reconstruction error between DAE and the proposed method.

According to the experiment results, we can know that through using the proposed method, the RMSE average of reconstruction error by 5-fold cross-validation decreased for all the data of three models with different skewed degrees both on AE and DAE, and it also speeds up the convergence for learning.

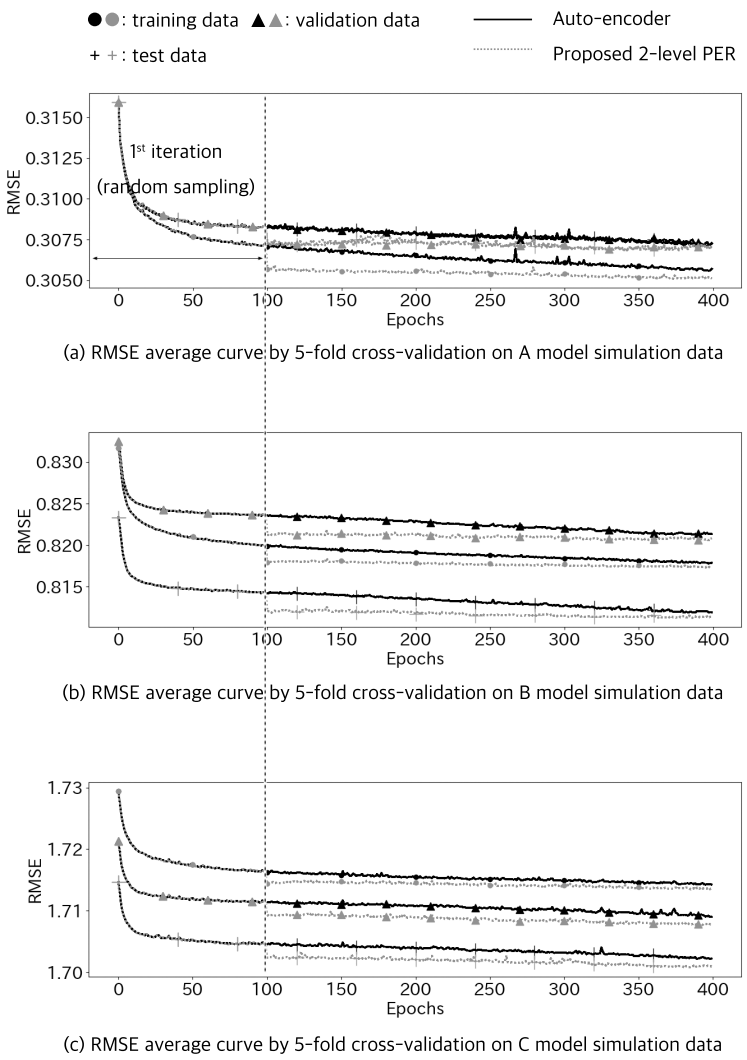


Figure 3.12: The RMSE average curve of AE and 2-level PER

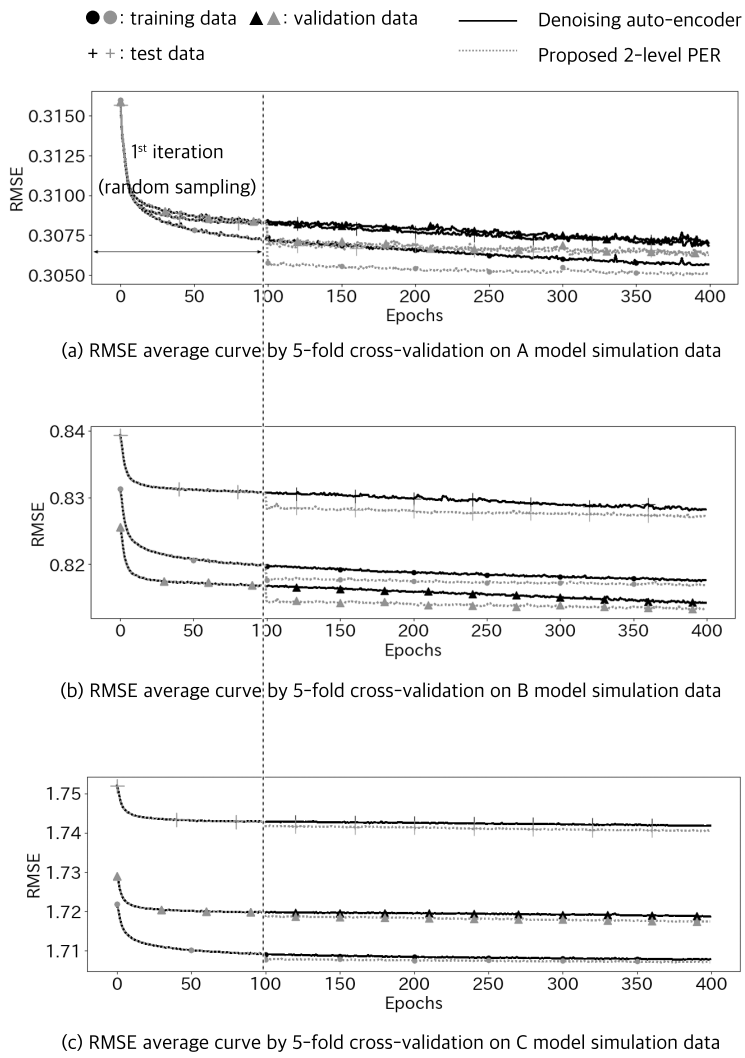


Figure 3.13: The RMSE average curve of DAE and 2-level PER

3.6 Summary

For the high dimension skewed data, a sampling method based on 2-level PER for AE and DAE is proposed. To evaluate the effectiveness, which without depending on a particular skewed degree, of this method, three models of high dimension skewed simulation data with a different skewed degree are generated.

Through appropriate parameter adjustments and evaluation experiments, the effectiveness of the proposed method,

1. Reducing the reconstruction error caused by the skewed data effectively,

2. Improving the generalization of the algorithm and the convergence speed of learning, have been confirmed.

In the next Chapter 4, we will use this method to improve accuracy for wave-making resistance estimation, when using deep neural network(DNN)[35] to perform a regression for ship shape data, the data is skewed and have bias on offset data, to improve the learning rate and accuracy of DNN, the input layer need some compressed features with generalization. The features can be extracted through this study to improve the accuracy of the regression.

Further, no only for this research, more usages of the other research are expected. and it is also expected to apply some practical problems in other research, and the proposed method will be applied to the actual high dimension skewed data, to improve the efficiency of feature extraction during learning, and further, verify the effectiveness of this method on a practical problem.

Of course, there are still many shortcomings in this method. For example, in the case of parameters, it is necessary to make adjustments through many preliminary experiments to make the proposed method get better performance, we treat this as a legacy topic in the current stage, and will consider a more intelligent optimization

solution, such as parameters adjustment automation based on the skewed degree of the dataset in the future.

Besides, when the dataset has a highly skewed degree, the original learner such as AE or DAE will get high reconstruction errors, we can reduce the high error caused because of skewed data, but can not fundamentally solve the high error problem of the learner. At the same time, for DAE, a learner that has some generalization performance itself, with the increase of the skewed degree and the high error problem of the learner, although the proposed method can achieve better results than the conventional methods, the performance is slightly reduced. These will be considered as the subject of future research.

4 Wave-making resistance estimation through deep learning considering the distribution of ship figure

4.1 Overview and the new proposed method

In Chapter 2 a wave-making resistance estimation method has been introduced. However, because of the bias of offset data, we proposed an improved feature extraction method for high dimension skewed data based on AE and PER in Chapter 3. And in this chapter, a wave-making resistance estimate method which considering the distribution of ship figures will be introduced.

The method in this chapter includes two phases,

1. First, improved auto-encoders, which reduce the dimension of the offset and the profile data at the same time considering the data distribution, are generated. While performing to the skewed offset data, use the method introduced in Chapter 3.

2. Subsequently, after the regularization of these data, a deep neural net for regression estimation of wave-making resistance is generated, this part is the same with Chapter 2.

With the consideration of data distribution, compared to the results in Chapter 2, the results of the experiment show that the proposed method in this chapter can improve the precision of wave-making resistance estimation furthermore. Figure 4.1 shows the new procedure which improved the procedure in Chapter 2 used 2-level PER sampling method in Chapter 3.

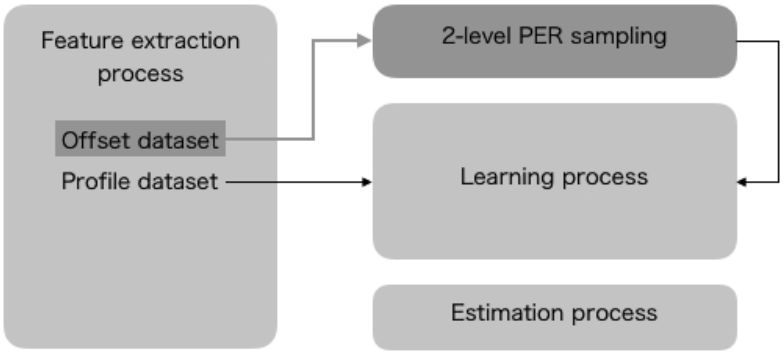


Figure 4.1: New procedure for offset sampling and estimating wave-making resistance from ship shape

4.2 Dataset

In this chapter, we used two types of dataset,

1. The dataset of 58 ships which is the same with the dataset in Chapter 2;
2. The increased dataset which include the shape parameters and water tank measurement data for other 587 ships, the definition of ship shape parameters is the same with the definition in Chapter 2.

When performing experiments in the next section, these two types of the dataset will be used, and comparison experiments will be performed.

4.3 Experiment

4.3.1 Experiment environment

The settings of DAE and DNN parameters are showed in Table 4.1.

About the adjustment of the α_1 and α_2, because of α is the weight for determining the selection probability and means the importance of the skewed degree of ship shape data, it depends on the dataset. Therefore to determine the α_1 and α_2 for each step of learning, we performed a grid search and the figure shows the result of reconstruction errors (RMSE) with α_1 and α_2 changes by heat map. Figure 4.2 shows the result, through the results, we set the values shown in Table 4.1.

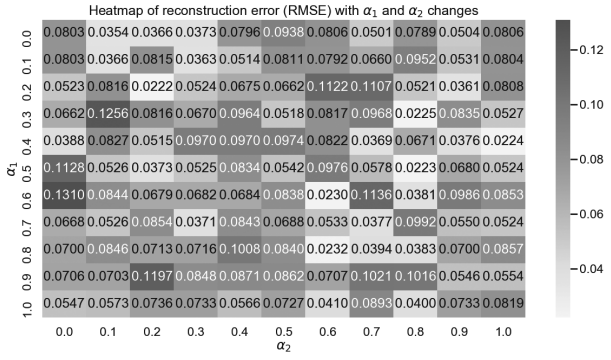


Figure 4.2: Heatmap of reconstruction error (RMSE) with α_1 and α_2 changes

And the spec of machine what we used in the experiment is,

1. CPU: Intel Xeon E3-1230

2. Memory: 16GB DDR-4 SDRAM

3. Graphics: NVIDIA Quadro P400 2GB

4. OS: Linux

Based on the above spec, both the method in Chapter 2 and proposed method in this chapter take 30 minutes for 58 ships, and 100 minutes for 587 ships on one cycle learning of DAE and DNN.

4.3.2 Experiment result

The result is, compare to using the method in Chapter 2 when using the proposed method of this chapter on processing 58 ships, the RMSE reconstruction error of offset data is reduced from 0.00501 to 0.00429. And the RMSE regression loss reduced from 0.078 to 0.047, it proves a 39.74% error reduction performance. Figure 4.4 shows the RMSE distribution of regression loss for sorted 58 ships. From the figure, we can know that with using the proposed method, regression loss reduced substantially and especially for the ship types which have a high regression error before.

Moreover, the reconstruction error reduced from 0.00601 to 0.00417 when processing 587 ships, and the regression loss reduced from 0.114 to 0.095, a 16.67% error reduction performance has been shown. Figure 4.6 shows the result of sorted 587 ships. Here, Because of the increase ship numbers, the types of the ship also increased, it made the skewed degree of dataset decrease, therefore compare to the 58 ships, the predominance of the proposed method decreased a little bit while processing 587 ships.

Even so, we can find that the hull form with high error is effective. To confirm the

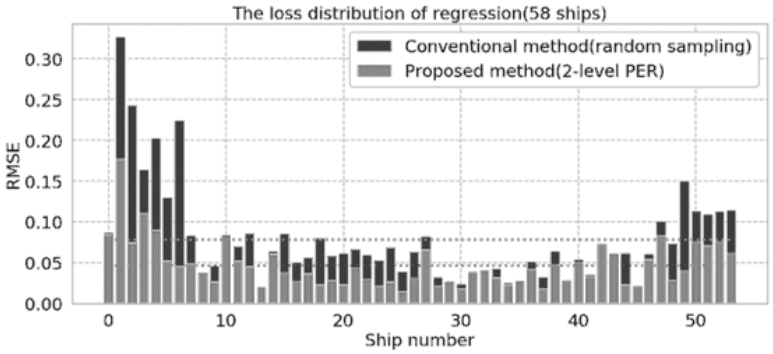


Figure 4.3: RMSE distribution of regression loss with 2-level PER sampling (in case 58 ships)

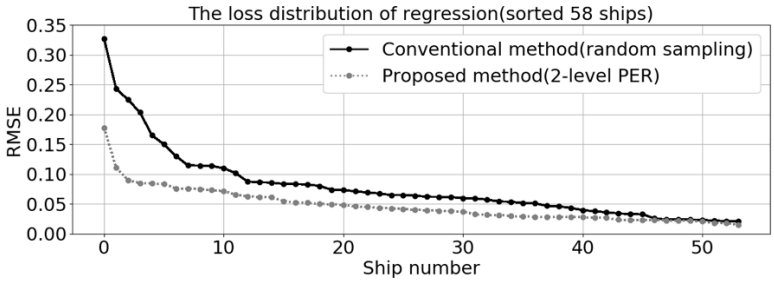


Figure 4.4: RMSE distribution of regression loss with 2-level PER sampling (in case sorted 58 ships)

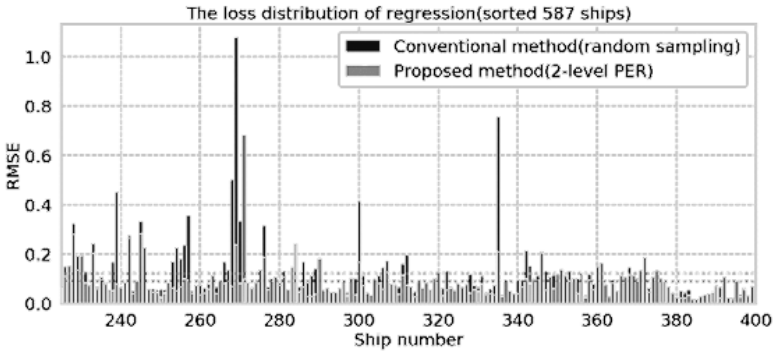


Figure 4.5: RMSE distribution of regression loss with 2-level PER sampling (a part of ships in case 587 ships)

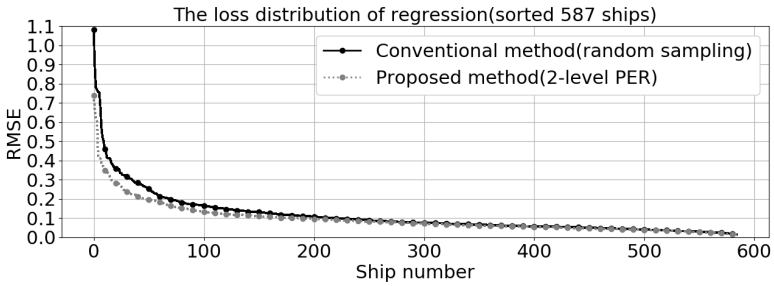


Figure 4.6: RMSE distribution of regression loss with 2-level PER sampling (in case sorted 587 ships)

effectiveness of the proposed method, the conventional method and the proposed method were compared on ship types with high regression error of the top 20%.

For samples with a high error of the top 20%, When the proposed method was used, the error was reduced by 22.9%. In the case of the remaining 80%, only 9.4% error reduction was achieved. Especially for samples with high error, the performance improved. It proved that when there are skewed data, the generalization performance can be maintained with the proposed method.

Figure 4.7 shows the comparison of top 20% and remaining 80%.

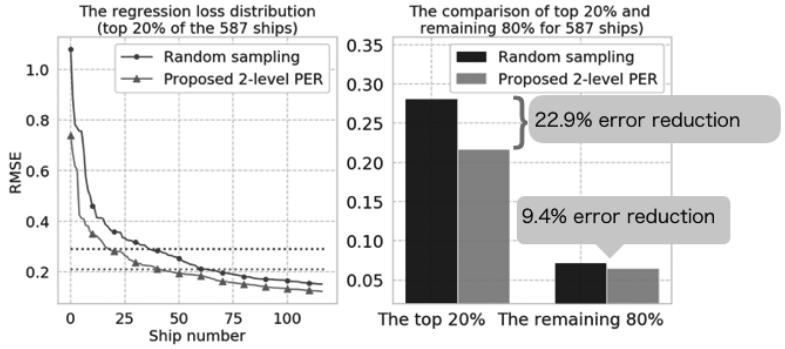


Figure 4.7: Comparison for top 20% and remaining 80% of 587 ships

Figure 4.8 shows the reconstruction error changes with epoch number increased. if not using sampling, the reduction of reconstruction error can not find nor training data, validation data, and test data. When using the proposed method which includes a sampling process, from 100 epoch which is the start point of sampling, reconstruction error is also starting to reduce.

Figure 4.9 shows the results of the F_n-C_w regression curve comparison between not using sampling and using the sampling method. When not using sampling, with F_n increased, the deviation of C_w also increased, and the hump parts of F_n-C_w curve are also not represented distinctly. Here, the hump case is a type of

change in the relationship of F_n-C_w that appears in some hull forms. For these hull forms, the trend of the changes in the increase of speed within a certain range of ship speeds can be reversed. An appearance is a small number of data from all ship data. When using the sampling method, from the figure we can know that the reproducibility improved a lot.

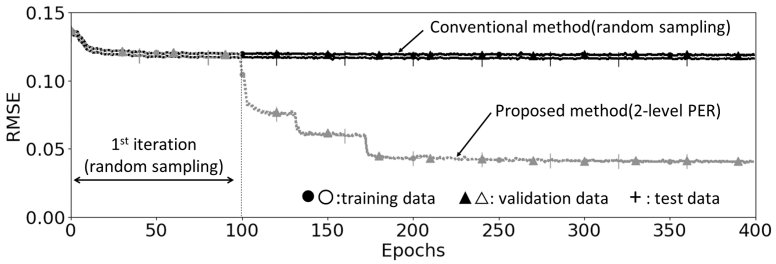


Figure 4.8: RMSE average curve by 10-fold cross-validation

4.4 Summary

For the distribution of the ships which have various shapes, based on the method in Chapter 2, a 2-level priority sampling based on the method in Chapter 3 has been used and improved the learning accuracy. With these methods, high-precision wave-making resistance estimation is realized.

As a consideration, if the combination of α is suitable for the data set, an appropriate training sample will be selected. However, if it deviates from the optimal α value by about 0.1, the error may increase. Here, if the step of α is made smaller, it is thought that the error changes smoothly, but it takes enormous time when using the grid search, and it becomes difficult to search for the optimal solution. Therefore, we are currently setting the increment of 0.1, but as a direction for further development, we can expect a smarter search method using Bayesian optimization.

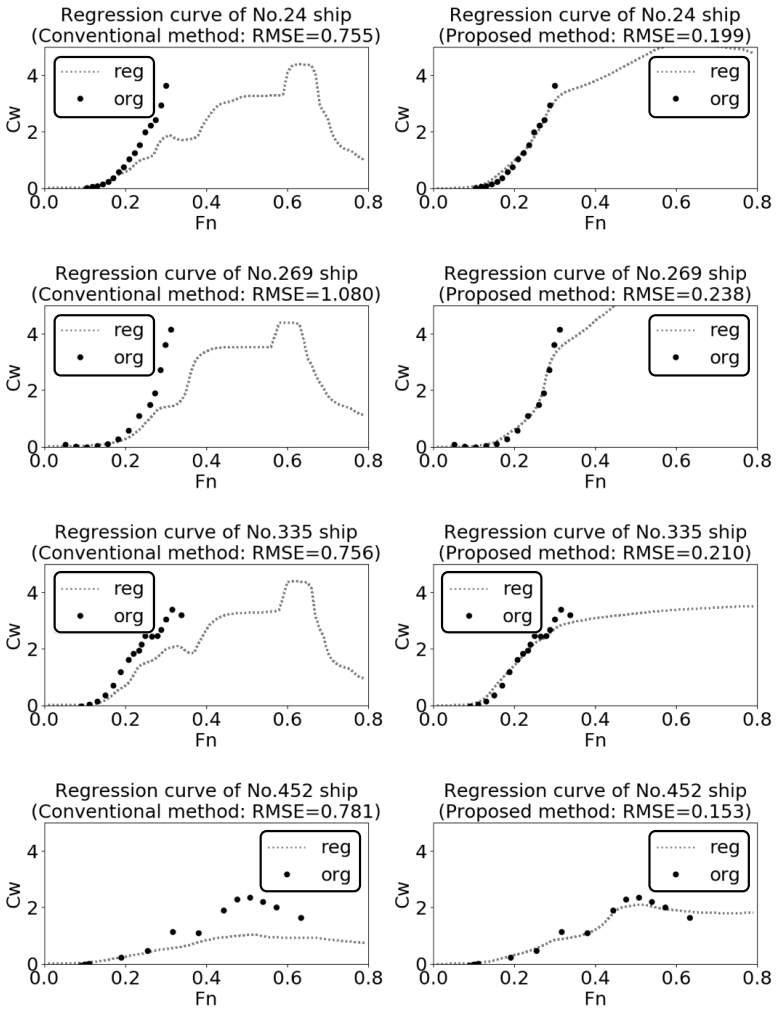


Figure 4.9: The examples of comparison results of the F_n-C_w regression. The left-column: conventional, the right-column: proposed method.

Table 4.1: Parameters of improved DAE and DNN for regression of

Profile number	1174×100
Offset number	22470×100
DAE sampling group stride	0.001
DAE sampling factor	$\beta = 0.5$
DAE sampling parameter	$\alpha_1 = 0.5, \alpha_2 = 0.8$
DAE sampling times	4
DAE epoch number for each sampling	$EP = 100$
DAE total epoch number	$400 = 4 \times 100$
DAE batch size	20
DAE layer	(100)-(64 tanh)-(32 tanh)-(4 tanh)-(32 tanh)-(64 sigmoid)-(100)
DAE Loss function	RMSE(Root Mean Squared Error)
DNN epoch number	10000
DNN batch number	250
DNN input number	$171 = 5 + 5 + 20 \times 2 \times 4 + 1$
DNN layer	(171)-(256 relu)-(128 relu)-(64 sigmoid)-(16 sigmoid)-(1)
DNN Loss function	MSE(Mean Squared Error)
DAE and DNN Optimizer	Adam

5 Conclusions and future work

5.1 Conclusions

In this dissertation, we proposed a high-precision estimation method through deep learning with considering data distribution. Meanwhile, throughout the theory of this study through a case in wave-making resistance estimation considering the distribution of ship figures for hull design support.

In Chapter 1, we raised the motivation of this study, clarified the significance of the study on energy saving and environmental protection through the case in wave-making resistance estimation of the shipbuilding industry. Moreover, we introduced the concept of wave-making resistance and the reason for performing a wave-making resistance estimation. Further, we explained that because of the inefficiency of existing estimation methods, it is necessary to propose an efficient estimation method such as end-to-end learning. Besides, we also introduced the feature extraction method that is expected to be used for the dimensionality reduction of ship shape data. The finally of the Chapter, we introduced the main idea of this study and gave an overview of the whole study.

In Chapter 2, we proposed a wave-making resistance estimation method from the shape of the ship through deep learning, which made end-to-end learning becomes possible on the non-uniform ship-shape data. The proposed method contains the pre-treatment method for non-uniformity data, the feature extraction solution through denoising auto-encoder and discrete Fourier transform, the regression pro-

cess through deep learning. At the same time, the experiment performed in this chapter evaluated that the proposed method can estimate wave-making resistance directly with high precision. At the end of this chapter, an issue about the learning bias due to the characteristic of skewed data has been discovered.

In Chapter 3, to solve the remaining issue which begot because of the skewed data, at the same time raise the generalization performance of the method in Chapter 2, we proposed an improved auto-encoder based on 2-level prioritized experience replay for high dimension skewed data. To evaluate the proposed method, three models of high dimension simulation dataset which on different skewed degrees are generated. To prove the universal validity of the method, the three models of simulation datasets have been used instead of the ship shape data. The experiment performed on both auto-encoder and denoising auto-encoder, and the results of evaluating experiments shown that the proposed method can get lower reconstruction error than the conventional method when the dataset is skewed.

In Chapter 4, the method in Chapter 3 has been applied on the ship shape data, and according to the method we improved the method in Chapter 2, proposed a wave-making resistance estimation method with considering the distribution of ship figure. Further, to evaluate the proposed method, based on the previous ship-shape dataset, a dataset with another larger dataset of ship shape is to be used. In the experiment section, a comparative experiment conducted on both two datasets, and the results of the experiment show that the proposed method in this chapter can improve the precision of wave-making resistance estimation furthermore.

It is expected that the method introduced in this study will be used more generally. The problem classes that can be considered when using the method generically are as follows. (1) Feature extraction problems for non-uniform dataset when the

data have smooth changes and the frequency components that can be extracted fall within a certain range; (2) Learning process optimization problems when the error of only a small part of the data is high, and it is considered to be a problem due to the bias of data; (3) Problems that require that explainable features obtained from learning and further applications are needed based on these features.

In the appendix of this dissertation, we give several application solutions through this study for hull design support, which should be a hint for the needs of ship industries. Based on the wave-making resistance estimator generated through this study, A hint of idea about using variational auto-encoder to perform part morphing, and usages of performing completed morphing for the ship shape through inverse discrete Fourier transform are raised in Appendix B. At the same time, with the necessity for the similarity evaluation between ships, because of the different length of the ships, it is recommended that using a dynamic solution such as dynamic time warping instead of a static one. We also raise an example that performing a ship-shape optimization in Appendix C, which based on the wave-making resistance estimator of this study, through using a part morphing by variational auto-encoder and real-coded genetic algorithm.

5.2 Future work

An issue which about the completed solution for hull design of shipbuilding has been raised, we will consider some solutions for this issue and try to solve it in the future work. For example, when performing the experiments for this study, we used some random factors. indeed, starting from random as initial parameters during learning may cause some uncertainty. Nevertheless, as the learning progresses, the model is developed in the direction of reducing the interference of random

factors. For models that have completed the learning, while ensuring a certain generalization ability, it applies to the determinism of the current world. We hope that this research can be applied in many fields, not just in the field of ship design, but more to solve growing and urgently needed problems. We also hope that this research can be put into use as soon as possible under the premise of ensuring quality and accepted the test of practical application, to find problems and improve the generalization ability of this research.

Appendix A

About the optimization problem of the ship shape for hull design support

Here, about how to support the ship hull optimization through the theory of this dissertation will be described, several examples of the applications about optimization method for ship hull shape design support through the estimator by this theory will be introduced. The target is the optimization method for ship shape according to the generated estimation model.

Hull design support has important industrial significance, in the shipbuilding industry, there is a big demand to optimize the shape of the hull. A high-precision wave-making resistance estimator through deep regression which considered the ship figure distribution already be generated through this dissertation. Because of the generated estimator has been evaluated in several experiments and proven to have an applicable precision accuracy, therefore we are considering to use it as an evaluation standard for the hull design.

About the hull design support, there are several methods. To illustrate how to use the generated estimator for optimization, The methods include morphing methods such as variational auto-encoder(VAE) or discrete Fourier transform(DFT), to generate a variation of ships, and metaheuristic methods such as genetic algorithm(GA) to perform an optimization will be introduced.

Appendix B

Hull morphing and variation generation

Part morphing through variational auto-encoder(VAE)

As what mentioned in Chapter 3, variational auto-encoder(VAE) is also one of the variants of AE, due to its unique variability because of the controllable latent space, it is often used as generation model, here, we also want to use VAE to generate the variation of ships. However, the length of the ships are not the same, therefore, it is impossible to use the offset data as input of VAE directly.

To solve this problem, because the hull part excluding fore and aft usually have little changes, it can be considered only apply the learning on the part of fore and aft. Further, because of the continuous offset does not change suddenly, we assign a fixed-length parameter k for both fore and aft.

Figure 1 shows the encoder part, Figure 2 shows the decoder part of VAE. Table 1 shows the specifications of VAE which using to generate fore and aft variation in the experiment.

About the loss function for VAE, KL Divergence is used Equation 1 shows the definition of the loss function, and Figure 3 shows the loss rate on training data and test data of VAE learning.

Figure 4 shows the aft generation result of VAE.

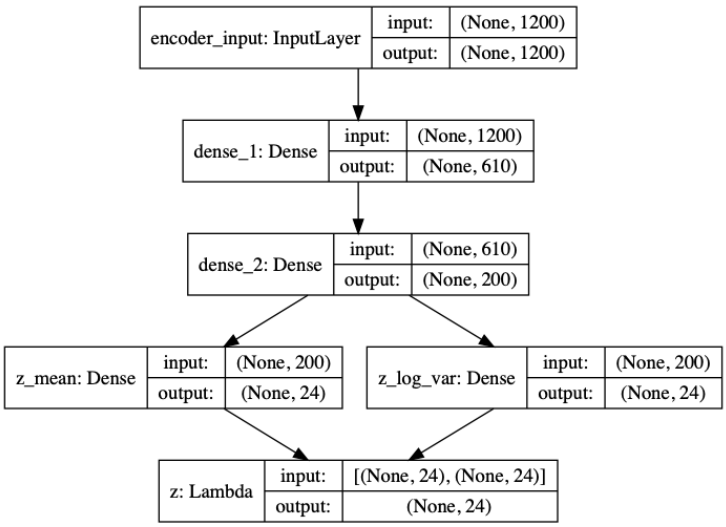


Figure 1: Encoder part of VAE, which compress fore and aft parameters to low-dimensional latent space

$$LOSS_{VAE} = MSE \times D_{input} - 0.5 \times [\sum_{k=i}^{0,max} (1 + z_{\sigma_i} - z_{\mu_i}^2 - e^{\sigma_i})] \tag{1}$$

Figure 5 shows the procedure for evaluation, this procedure should be used to support a hull design.

Completed morphing through inverse discrete Fourier transform(IDFT)

VAE can well performed on the fore and aft part of the hull, however, to reduce the amount of calculation for the integration on the each parts of the hull, a completed morphing for the hull is expected.

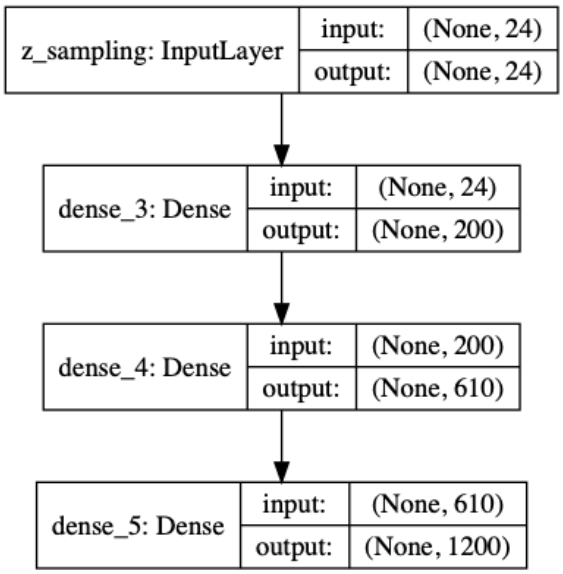


Figure 2: Decoder part of VAE, which generate variated fore and aft parameters from latent space

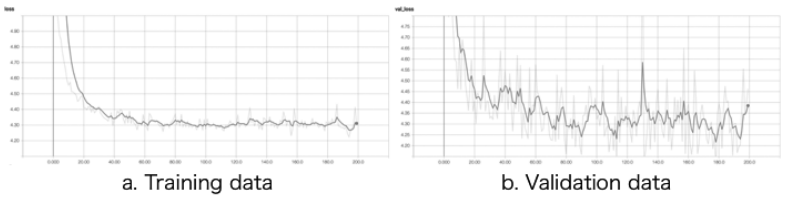


Figure 3: The learning loss per epochs of VAE

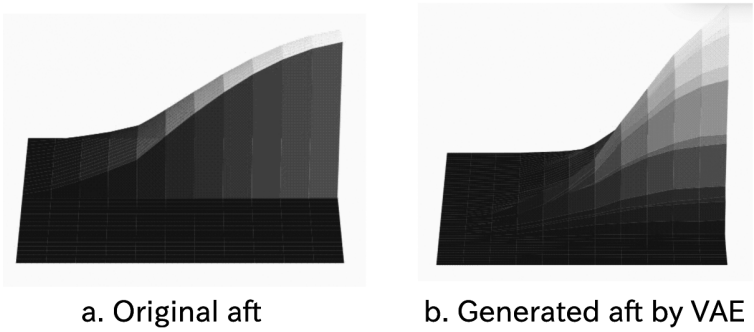


Figure 4: An aft generation example of the VAE morphing result

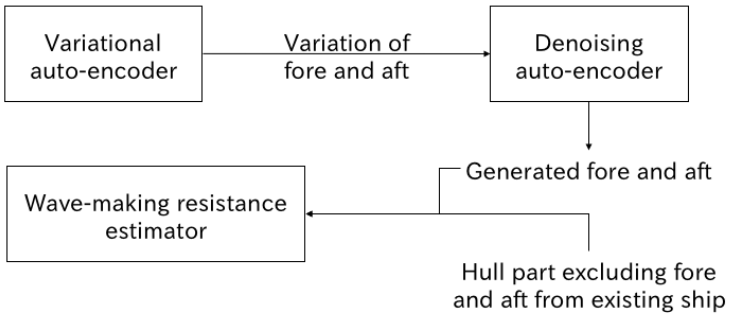


Figure 5: Procedure of evaluating the ship shape by wave-making resistance estimator

Table 1: Parameters of VAE for generating fore and aft variation

Offset number	$587 \times k \times 100$
Fore and aft length parameter	$k = 12$
Data extension	587×30
VAE input	587×1200
VAE epoch number	$EP = 500$
VAE batch size	20
VAE layer	(1200)-(610 tanh)-(200 tanh)-(24 tanh)-(200 tanh)-(610 tanh)-(1200)
Loss function	MSE, KL Divergence

From Chapter 2, we have already known that the offset data can be approximated represent as fixed low-dimension components of frequency, and we generated the fixed-length offset features through the discrete Fourier transform(DFT). Here, according to a similar consideration, it is logically reasonable that recovery of the low-dimensional representation of offset data from the low-dimension components of frequency through inverse discrete Fourier transform(IDFT)[36]. Figure 6 shows the procedure for generating a completed offset feature from the frequency component of two ships.

Through this procedure, there are two main steps,

1. Firstly, prepare the offset features of two ships which performed interpolation and zero-padding;

2. Subsequently, generate variated frequency component through average calculation and recovery it to a new offset feature.

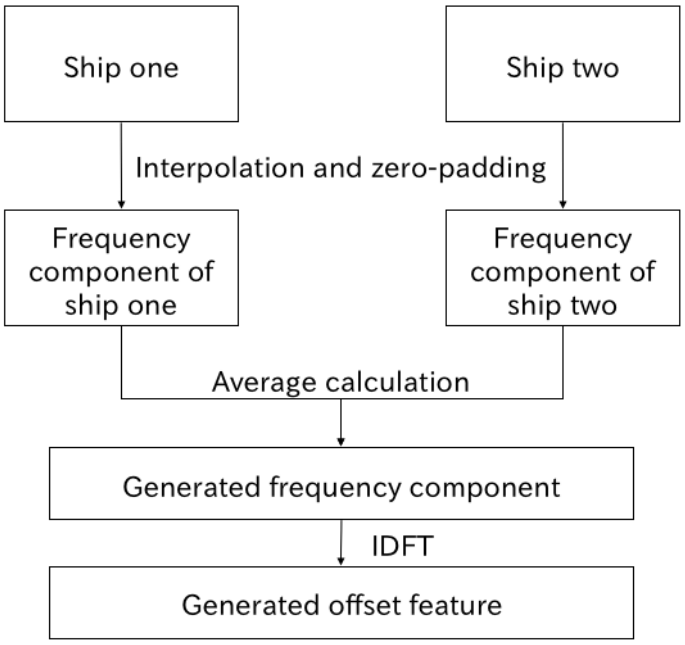
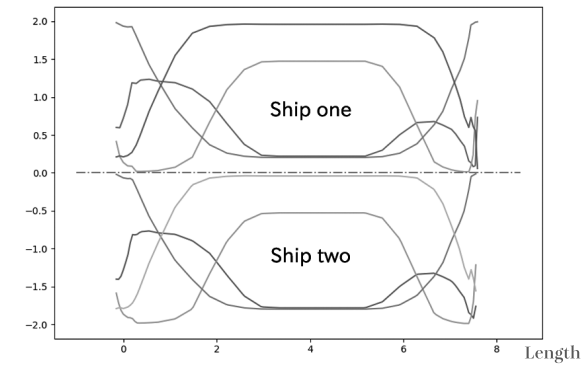
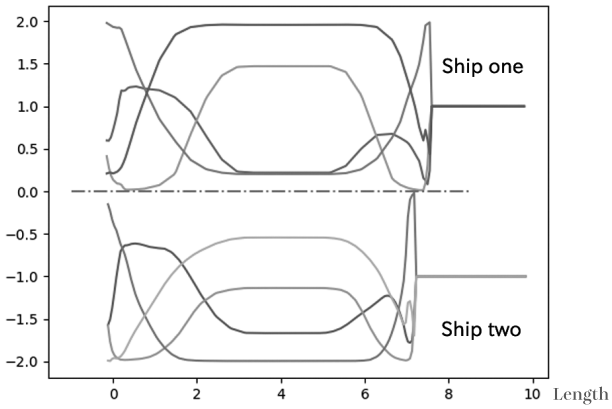


Figure 6: Procedure of completed feature generation from frequency component of two ships

Figure 7 shows the interpolated offset features of ship one and ship two, and their zero-padding performed presentation, here, the coordinates of the y-axis can be ignored because for the convenience of comparison, the relative movement up and down for ship one and two has been performed.



a. The interpolated offset features of ship one(the part above the dash-dot line) and ship two(the part below the dash-dot line)



b. The zero-padding examples for offset features of ship one(the part above the dash-dot line) and ship two(the part below the dash-dot line)

Figure 7: The interpolated and zero-padding performed offset features of ship one and ship two

Figure 8 shows the variated frequency component generation.

Within the above results, a completed morphing for a hull becomes achievable through the variation generation procedure which is shown in Figure 9.

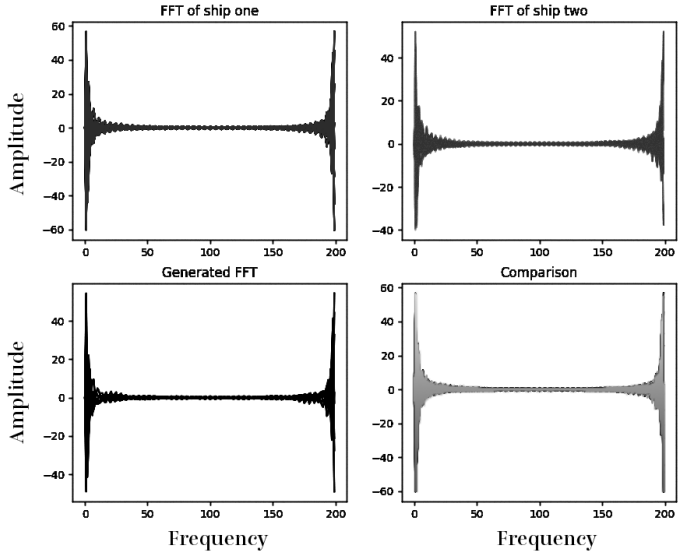


Figure 8: The variated frequency component generation from two ships

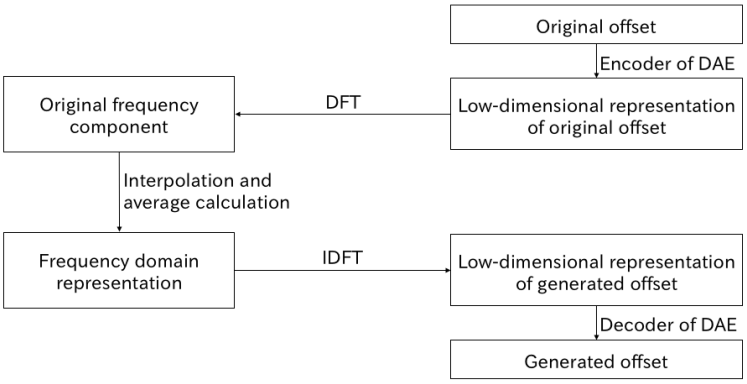


Figure 9: Procedure for a completed variation generation based on frequency domain representation adjustment

However, when to perform IDFT to generate offset feature, we found the generated features are unsatisfactory comparing to the generated features of by VAE, the reason for leading to the new shape of the ship be not reality is considered as follow,

1. Although DFT can retain a lot of features, the operation of IDFT is equivalent to accumulating losses multiple times, resulting in a cumulative error that reduces accuracy;

2. The two ships which used as input of the procedure have the low similarity, this also causes a large bias in the frequency domain, and make the cumulative error of IDFT increase.

To raise the feasibility of this method, it is necessary to evaluate the similarity of different ships.

Similarity evaluation through dynamic time warping(DTW)

In this study, the ships have different length, to evaluate the similarity of the different ships, we can not perform a measurement point by point, it is necessary to use a dynamic method on the length direction of the ship. Dynamic time warping(DTW)[37] is an algorithm that measures the degree of similarity between two signal sequences that differ in time or speed.

When performing DTW, we decompose each dimension of the offset features into one sequence and perform DTW processing one sequence by one sequence. To verify the validity of this method, 6 ships have been selected randomly, and DTW will be performed on every two ships. Figure 10, 11, 12 show the results of the DTW which performed on the 6 ships. Every figure shows the shortest distance route of each two ships, and the DTW distance of each two ships.

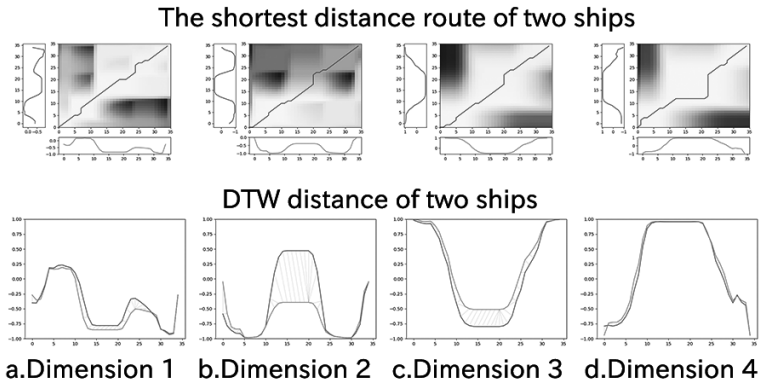


Figure 10: The shortest distance route and DTW distance of group A(ship 1, ship 2)

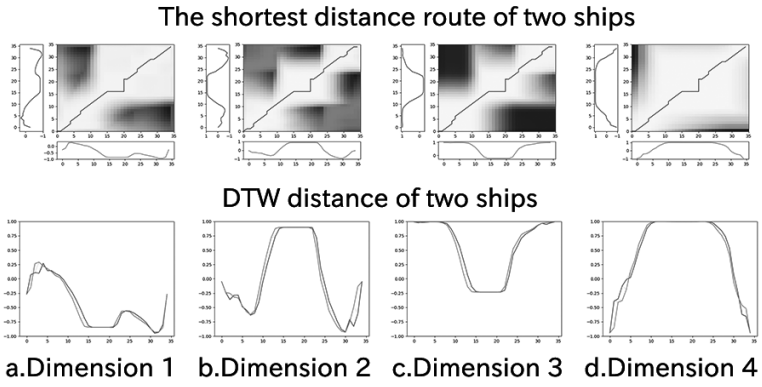


Figure 11: The shortest distance route and DTW distance of group B(ship 3, ship 4)

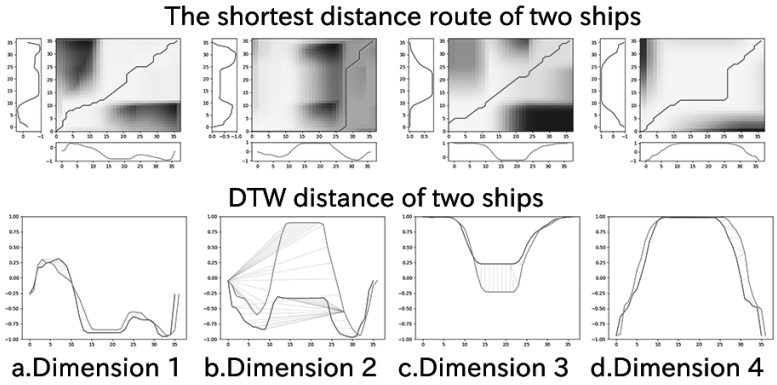


Figure 12: The shortest distance route and DTW distance of group C(ship 5, ship 6)

Through these figures, we found that although the variation of dimension 2 for each ship always be large, the DTW path draw by gray lines shown on DTW distance graph able to mapping accurately except the group C. We will make this as a future issue.

In the current step, we used the good case to different length offset features, and we can get the dynamic mapping path which shows in Figure 13. and each graph in this figure means the different dimensions of the two ships.

Using the DTW mapping path which we get, we generated an example of a new ship through the mapping path. Figure 14 shows the result.

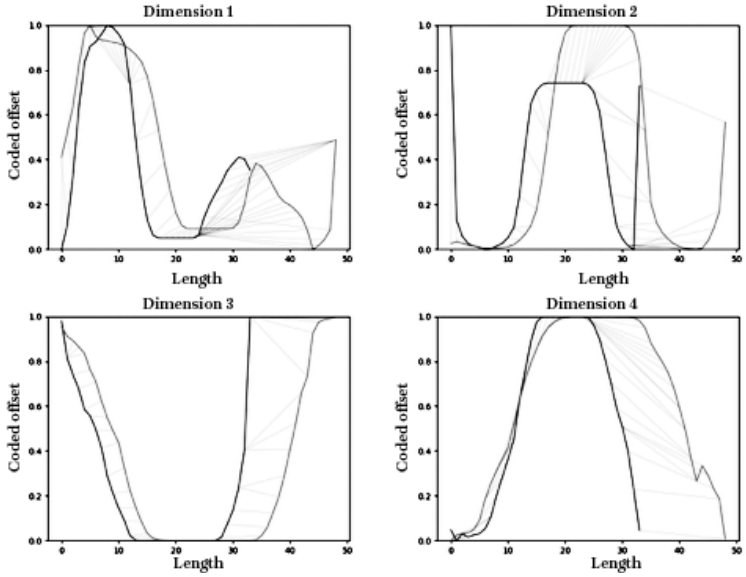


Figure 13: The dynamic mapping through DTW

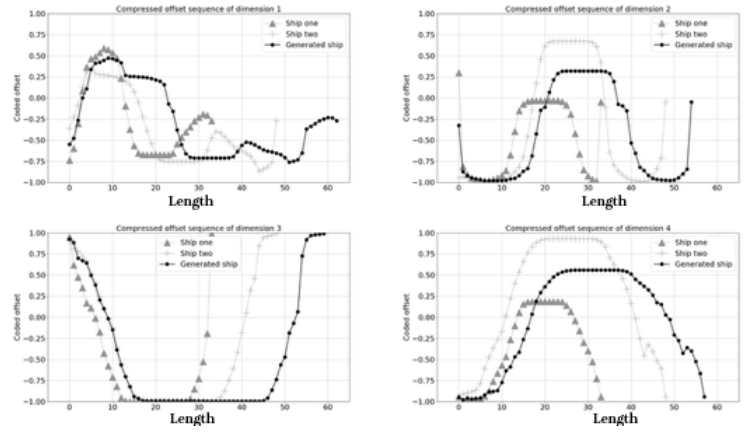


Figure 14: The new feature generated by DTW from two ships

Appendix C

Hull design support

Using denoising auto-encoder(DAE) to get low-dimension features of the ship, and then change the features little by little to generate variation of ships, the estimator can be used to evaluate if the new shape of the ship is better than existing ships. Further, considering the solution for the optimization process which should give support on hull design.

Based on the generated wave-making resistance estimator through this study, there are many possibilities of solutions for hull design, Here, an example as a hint for hull design will be raised.

Genetic algorithm(GA)[38] is a classic algorithm which as a representative meta-heuristic method[39], has been widely recognized as an optimization method. It can reply to the demands of industrial applications, and respond to multi-objective optimization problems, it also has a simplicity of the framework and high search performance especially in optimization problems with high-dimensional multidimensional variables. Based on these characteristics of GA, hence it is reasonable to apply the method on the hull shape optimization problem. Real-coded genetic algorithm(RCGA)[40] is a branch of GA, which using continuous parameters as they are without modifying the genes to the bit string. In the experiment below, we will use the fore and aft variation results of VAE which got in Section 5.2 as the first generation of RCGA, and use the wave-making resistance estimator to

evaluate every generation.

Figure 15 shows the procedure the hull optimization used the fore and aft variation results of VAE through RCGA and wave-making resistance estimator.

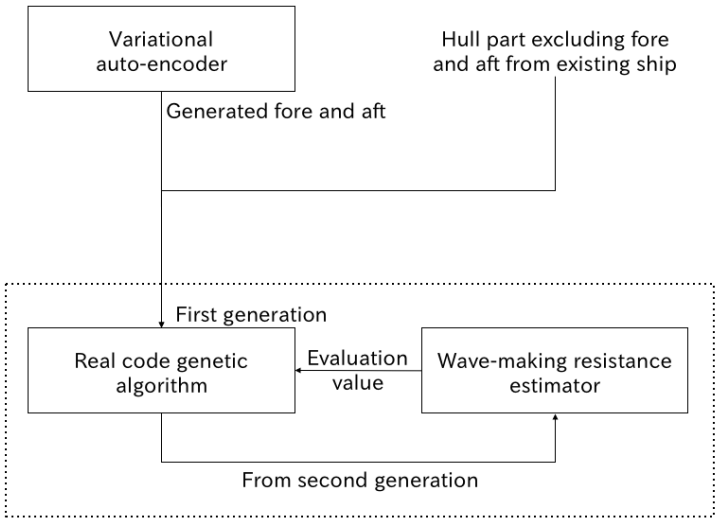


Figure 15: Procedure of the hull optimization through RCGA and wave-making resistance estimator

Evaluation value of RCGA

To evaluate a ship shape if it is an optimized hull need considering many factors, in light of the evidence, the most important is the wave-making resistance curve which we can use the estimator to generate from any ship shape. However, the production of ships always needs some of the other constraints, for example, ships with different main speeds, as well as ships with different cargo requirements, need to consider the different range of wave-making resistance. Here we only raise an example of how to use this study to support ship production. In the experiment,

we let RCGA to evaluate a cumulative C_w range of the ship when $0.01 \leq F_n \leq 0.8$, and the measurement of the F_n stride is set to $\Delta F_n = 0.001$. Figure 16 shows the evaluation value what RCGA used in the experiment, and Equation 2 shows the calculation for the evaluation value $Eval$ of RCGA.

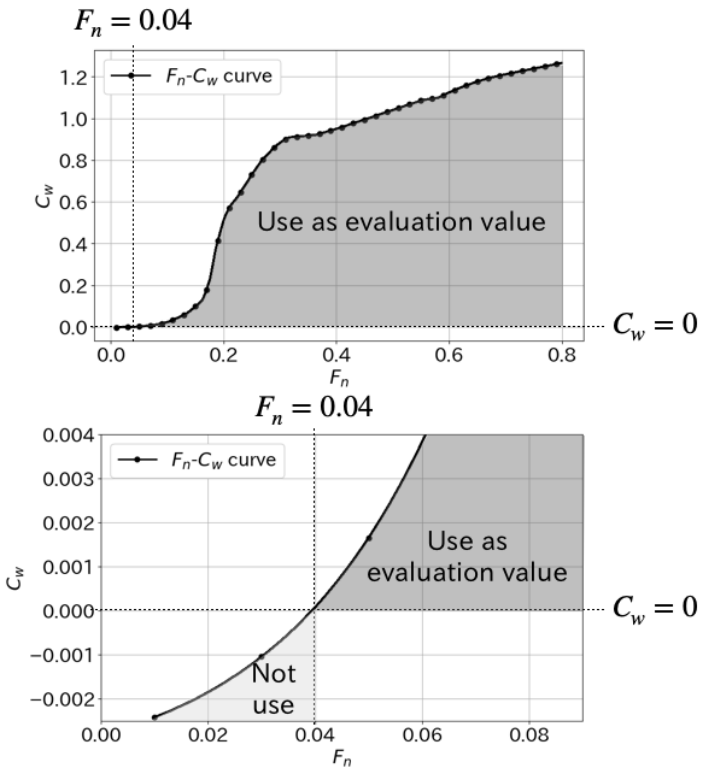


Figure 16: The evaluation value what RCGA used in the experiment

$$Eval = \sum_{i=0.01}^{0.8} [C_w(i) \times \Delta F_n] \tag{2}$$

Crossover and mutation of RCGA

Crossover and mutation are two basic operators of GA. In the experiment, we will use simulated binary bounded(SBX) which is a crossover method and polynomially bounded which is a mutation method for RCGA.

1. Simulated binary bounded(SBX)

   SBX is a crossover method for RCGA, and the parameter η control the variation between parents and children. Large variation from a small η and small variation from a large η value. The generation of two children from two parents is according to the following polynomial probability distribution.

   $$C(\beta) = 0.5(\eta_c + 1)\beta^{\eta_c}, \beta \leq 1 \tag{3}$$

   $$C(\beta) = 0.5(\eta_c + 1)\frac{1}{\beta^{\eta_c+2}}, \beta > 1 \tag{4}$$

   Here, η_c is a non-negative distribution parameter, let parents are $x_1, x_2(x_1 < x_2)$, then their children will be $c_1, c_2(x_l \leq c_i \leq x_u, i = 1, 2)$. therefore, the process for generation of two children is,

   (a) Generate uniform random number u range on $0 \leq u \leq 1$;

   (b) Calculate $\beta_{q,i}(i = 1, 2)$ according to probability distribution $C(\beta)$;

   (c) Calculate c_1, c_2 and generate the children.

2. Polynomial bounded(PB)

   PB is a mutation method for RCGA. It also has a parameter η to control the variation between parent and child which is similar to SBX, and the relationship between variation and η is the same with SBX.

Experiment of RCGA

Table 2 shows the specifications of RCGA.

Table 2: Parameters of RCGA for hull optimization

Generations	30
Individuals	30
Selection	Tournament selection
Crossover rate	0.6
Mutation rate	0.3

Figure 17 shows the result for fitness curve of RCGA.

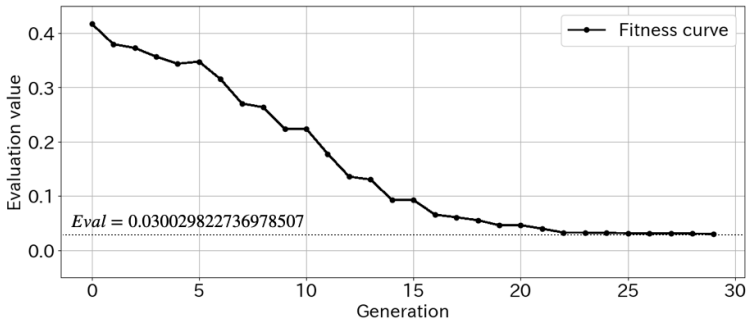


Figure 17: The fitness curve of RCGA

With the generation number increase, the wave-making resistance evaluation value can decrease well, however, in the real production for hull shape, the last generation can not be used directly. That is because

1. The actual industrial application must be considering more of the limiting factors and add them to the constraints of RCGA or other metaheuristic methods.

2. This experiment is just a hint for hull design support, therefore it used the common crossover and mutation method of RCGA, however, to raise the

availability, the crossover and mutation method for the exclusive use of the ship shape must be proposed, this could be an issue in the future.

Bibliography

[1] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[3] Bao-ji Zhang, Kun Ma, and Zhuo-shang Ji. The optimization of the hull form with the minimum wave making resistance based on rankine source method. *Journal of Hydrodynamics*, 21(2):277–284, 2009.

[4] 財務省統計局普通貿易統計. *http://www.customs.go.jp/toukei/info/*, 2017.

[5] John V Wehausen and Edmund V Laitone. Surface waves. In *Fluid Dynamics/Strömungsmechanik*, pages 446–778. Springer, 1960.

[6] JC Lewthwaite, AF Molland, and KW Thomas. An investigation into the variation of ship skin frictional resistance with fouling. *Royal Institution of Naval Architects Transactions*, 127, 1985.

[7] Roger Brard. Viscosity, wake, and ship waves. *Journal of Ship Research*, 14(4), 1970.

[8] Jerome H Saltzer, David P Reed, and David D Clark. End-to-end arguments in system design. *ACM Transactions on Computer Systems (TOCS)*, 2(4):277–288, 1984.

[9] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71):13, 2009.

[10] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti A Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2008.

[11] Ian Jolliffe. *Principal component analysis*. Springer, 2011.

[12] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[13] Jacek M Zurada. *Introduction to artificial neural systems*, volume 8. West publishing company St. Paul, 1992.

[14] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

[15] Philip E Gill, Walter Murray, and Margaret H Wright. Practical optimization. 1981.

[16] A Cochocki and Rolf Unbehauen. *Neural networks for optimization and signal processing*. John Wiley & Sons, Inc., 1993.

[17] Hajime Maruo, Kazuko Kasahara, Kazuo Suzuki, and Takehiko Kawamura. On the ship form of minimum wave resistance with a bow bulb. *Journal of the Society of Naval Architects of Japan*, 1975(138):1–11, 1975.

[18] 野澤和男 and 児玉良明. Cfd による船体流場の推定と設計への応用. *Techno marine* 日本造船学会誌, 806:542–551, 1996.

[19] 日野孝則. 船舶の作る波の cfd 解析. **数理解析研究所講究録**, 1989:8–15, 2016.

[20] Björn Tings, Carlos Augusto Bentes da Silva, and Susanne Lehner. Dynamically adapted ship parameter estimation using terrasar-x images. *International Journal of Remote Sensing*, 37(9):1990–2015, 2016.

[21] Anthony F Molland. *The maritime engineering reference book: a guide to ship design, construction and operation*. Elsevier, 2011.

[22] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408, 2010.

[23] Raghu Machiraju and Roni Yagel. Reconstruction error characterization and control: A sampling theory approach. *IEEE Transactions on Visualization and Computer Graphics*, 2(4):364–378, 1996.

[24] Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. 1998.

[25] S Surendran and J Venkata Ramana Reddy. Numerical simulation of ship stability for dynamic environment. *Ocean Engineering*, 30(10):1305–1317, 2003.

[26] 立花誠人 and 村田剛志. 構造特徴とグラフ畳み込みを用いたネットワークの半教師あり学習. **人工知能学会論文誌**, 34(5):B–IC2_1, 2019.

[27] Horace B Barlow. Unsupervised learning. *Neural computation*, 1(3):295–311, 1989.

[28] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[29] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2(5):6, 2017.

[30] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

[31] Mario Szegedy. The dlt priority sampling is essentially optimal. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 150–158. ACM, 2006.

[32] François Chollet et al. Keras: The python deep learning library. *Astrophysics Source Code Library*, 2018.

[33] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[34] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[35] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[36] David J Young and Norman C Beaulieu. The generation of correlated rayleigh random variates by inverse discrete fourier transform. *IEEE transactions on Communications*, 48(7):1114–1127, 2000.

[37] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[38] Darrell Whitley. A genetic algorithm tutorial. *Statistics and computing*, 4(2):65–85, 1994.

[39] Fred W Glover and Gary A Kochenberger. *Handbook of metaheuristics*, volume 57. Springer Science & Business Media, 2006.

[40] Alden H Wright. Genetic algorithms for real parameter optimization. In *Foundations of genetic algorithms*, volume 1, pages 205–218. Elsevier, 1991.

Publication

Reviewed journal papers

- 李 欣, 新井 洋, and 濱上 知樹. 船型データの分布を考慮した深層学習による造波抵抗推定. 電気学会論文誌 C（電子・情報・システム部門誌）, 140(3):391–397, 2020.

- Xin Li and Tomoki Hamagami. An improved auto-encoder based on 2-level prioritized experience replay for high dimension skewed data. *Adaptation, Learning and Optimization*, 12(1):93–105, 2020. Springer.

Reviewed international conference papers

- Xin Li, Masaya Nakata, Hamatsu Fumiya, and Tomoki Hamagami. A research about anomaly detection method for multidimensional time series data. In *The International Conference on Electrical Engineering*, 2017.

- Xin Li and Tomoki Hamagami. An improved auto-encoder based on 2-level prioritized experience replay for high dimension skewed data. In *Proceedings of the 23rd Asia Pacific Symposium on Intelligent and Evolutionary Systems*, 2019.

Domestic conference papers

- Xin Li and Tomoki Hamagami. Prioritized sampling method for autoencoder to reduce loss rate for skewed data. In *The 28th Fuzzy, Artificial Intelligence, Neural Networks and Computational Intelligence*, 2018.